Acquisition and recognition of natural landmarks for vision-based autonomous robot navigation

Livatino, Salvatore

Laboratory of Computer Vision and Media Technology

Ph.D. Dissertation

Acquisition and Recognition of Natural Landmarks for Vision-Based Autonomous Robot Navigation

Salvatore Livatino

Department of Health Science and Technology

Aalborg University 2003

AUTHOR

SALVATORE LIVATINO

Salvatore Livatino studied Computer Science at the University of Pisa, Italy, (MSc.1993), with a specialization in Robotics, Computer Vision and Computer Graphics undertaken at the Scuola Superiore S.Anna Pisa, where he pursued his research activity in the coming years at the PERCRO and ARTS Lab (1992-95). Then, visiting researcher at the School of Computer Studies, University of Leeds, UK, (1995), and at the MOVI Lab, INRIA Grenoble, France, (1996).

In 1997 he joined the LIA/CVMT Lab at Aalborg University, Denmark, where his research activity proceeded primarily in the field of vision-based robot navigation and realistic synthesis of virtual views. This lead to the completion of his PhD in 2003. During this time at CVMT, Salvatore Livatino was also involved in various EU projects, including the design of project proposals for the EU FP5/6, in teaching activities as well as collaborating with the School of Informatics at the University of Edinburgh, UK, (2001), in the research area of 3D model reconstruction from analysis of range images.

Salvatore Livatino's main publications concern: autonomous robot navigation based on the use of natural landmarks, ([10], [120], [93], [92], [94], [88]); realistic image synthesis, ([91], [14], [89], [90]); 3D model reconstruction ([24], [23]).

Acquisition and Recognition of Natural Landmarks for Vision-Based Autonomous Robot Navigation

A Ph.D. dissertation
by

Salvatore Livatino

Laboratory of Computer Vision and Media Technology
Department of Health Science and Technology
Aalborg University, Denmark
E-mail: sl@cvmt.dk
URL: http://www.cvmt.dk/~sl

August 2003

This dissertation was submitted in May 2003 to the Faculty of Engineering and Science, Aalborg University, Denmark, in partial fulfillment of the requirements for the Doctor of Philosophy degree.

While the first edition was approved, this second edition includes revisions in accordance with comments from the adjudication committee.

The following adjudication committee was appointed to evaluate the thesis:

Professor Giovanni A. Muscato, Ph.D.
Dipartimento Elettrico, Elettronico e Sistemistico (DEES)
Faculty of Engineering
University of Catania
Catania, Italy

Professor Robert B. Fisher, Ph.D.
Institute of Perception, Action and Behaviour (IPAB)
School of Informatics
University of Edinburgh
Edinburgh, United Kingdom

Professor Erik Granum, Ph.D. (committee chairman)
Computer Vision and Media Technology Laboratory
Department of Health Science and Technology
Aalborg University
Aalborg, Denmark

Supervisor
Associate Professor Claus B. Madsen, Ph.D.
(member of the committee in a non-voting capacity)
Computer Vision and Media Technology Laboratory
Department of Health Science and Technology
Aalborg University
Aalborg, Denmark

Abstract

The use of landmarks for robot navigation is a popular alternative to having a geometrical model of the environment through which to navigate and monitor self-localization. If the landmarks are defined as special visual structures already in the environment then we have the possibility of fully autonomous navigation and self-localization using automatically selected landmarks.

The thesis investigates autonomous robot navigation and proposes a new method which benefits from the potential of the visual sensor to provide accuracy and reliability to the navigation process while relying on naturally available environment features (natural landmarks). The goal is also to integrate techniques and algorithms (also related to other research field) in the same navigation system, in order to improve localization performance and system autonomy.

The proposed localization strategy is based on a continuous update of the estimated robot position while the robot is moving. In particular, using a triangulation algorithm based on three landmarks the robot position can be estimated each time three landmarks are observed. If the selected landmark triplet is an optimal triplet, the estimated robot position is accurate and errors do not accumulate.

In order to make the system autonomous, both acquisition and observation of landmarks have to be carried out automatically. The thesis consequently proposes a method for learning and navigation of a working environment and it explores automatic acquisition and recognition of visual landmarks. In particular, a two-phase procedure is proposed: first phase is for an automatic acquisition of visual-landmarks, second phase is for estimating robot position during navigation (based on the acquired landmarks).

Automatic acquisition of landmarks emphasizes map-building. In particular, feature extraction and their mapping inside a workspace, raising the question of how this should be best achieved to provide reliable localization information. The thesis proposes a set of strategies based on panoramic acquisition, attention selection, and stereo reconstruction. When properly managed these allow building a suitable subset of landmarks to be used for self-localization.

Since visual landmarks acquired during the first phase have to be recognized during the second phase, a method is proposed for automatic recognition of self-learned landmarks. The aim is to make the recognition tolerant to positional errors and changes in viewpoint by a new method which better exploits the possibility offered by the proposed learning method, based on techniques adopted from the field of Realistic Virtual View Synthesis.

The feasibility and applicability of the proposed method is based on a system with a simple setup. The novelty and potentiality, are in combining algorithms for panoramic view-synthesis, attention selection, stereo reconstruction, triangulation, optimal triplet selection, and image-based rendering.

The system has been tested to evaluate its basic structure in order to achieve sufficiently reliable performance under realistic conditions. Experiments are presented which demonstrate that the system can automatically learn and store visual landmarks, and later recognize these landmarks from arbitrary positions and thus estimate robot position and heading.

Resume

"Indlæring og genkendelse af landemærker til vision-baseret robot navigation"

Brugen af landemærker til robot navigation er et populært alternativ til en komplet geometrisk model i forbindelse med positionsbestemmelse af en robot der navigerer rundt. Hvis landemærkerne er definerede som eksisterende visuelle strukturer i omgivelserne bliver det muligt at opnå komplet autonomisk navigation og positionsbestemmelse baseret pa automatisk udvælgelse af landemærker.

Denne afhandling undersøger autonom robot navigation og foreslår en ny metode som udnytter fordelene ved en vision-baseret sensor til at gøre navigation robust og nøjagtig ved at bruge naturligt forekommende landemærker. Målet er også at integrere teknikker og algoritmer (inklusive nogle fra andre forskningsområder) i eet navigationssystem, med henblik på at opnå forbedret robotlokalisering og system autonomi.

Den foreslaæde lokaliseringsstrategi er baseret på en kontinuert opdatering af robottens position mens den bevæger sig. En trianguleringsmetode baseret på tre landemærker bliver brugt til at beregne robottens position hver gang tre landemærker er observerede. Hvis de tre valgte landemærker er optimale bliver den beregnede position nøjagtig og fejl akkumulerer ikke.

For at gøre systemet autonomt skal både indlæring og observation af landemærker gøres automatisk. Afhandlingen foreslår derfor en procedure med to trin: første trin laver automatisk indlæring af visuelle landemærker, og andet trin laver positionsbestemmelse under navigation, baseret på de under første trin indlærte landemærker.

Den automatiske indlæring indebærer en model-opbygning. Afhandlingen foreslår et sæt strategier baseret på en kombination af panoramisk billedoptagelse, udvælgelse af diskriminante visuelle kendetegn og stereo rekonstruktion. Tilsammen gør disse mekanismer det muligt automatisk at opbygge en model med landemærker der er egnede til brug for positionsbestemmelse.

Da visuelle landemærker indlært under første trin skal genkendes under andet trin udvikles en metode til automatisk genkendelse af indlærte landemærker. Målet er at gøre genkendelsen tolerant overfor positionsfejl og ændringer i synsvinkel. Den nye metode som udnytter mulighederne i det indlærte data er baseret på teknikker adopterede fra forskingsområdet vedrørende realistisk visualisering.

Det foreslaæde system udmærker sig ved at have en simpel opbygning, og det bibringer nye metoder indenfor kombinationen af panoramisk billedoptagelse, detektion af diskriminante kendetegn, stereo rekonstruktion, triangulering, optimal landemærke-udvælgelse, og billed-baseret visualisering.

Systemet er blevet testet for at demonstrere dets funktionalitet under realistiske forhold. Eksperimenter præsenteres som viser at systemet kan indlære visuelle landemærker, og senere bruge disse landemærker fra vilkårlige positioner til at beregne robottens position og retning.

Preface

This thesis describes the primary research activity I have been involved with in the period 1997 - 2003. In November 1996 I arrived in Aalborg, Denmark, being the winner of a 1 year scholarship for research specialization abroad (Perfezionamento all'Estero) funded by the Scuola Superiore S.Anna of Pisa, Italy. Thanks to the stimulating and supportive research environment I found at the Laboratory of Image Analysis (to become the Computer Vision and Media Technology Laboratory - CVMT), my research activity proceeded for many more years primarily on the field of Vision-Based Autonomous Robot Navigation, for most of the time funded by the EU TMR-project VIRGO (Vision-Based Robot Navigation Research Network, [112]). I have also worked as a research assistant being involved in various laboratory activities including the design of project proposals for the EU-FP5/6.

I would like to thank Erik Granum, who has assisted me in various ways throughout the entire period I have been in Aalborg, by making an effort in establishing the funding for my research activity, and for taking the time to assist me, encourage me and providing inspiring thoughts.

I would like to thank Claus Madsen for being a patient and educative supervisor and for his friendly and close scientific support full of useful advises.

I would like to thank the following people who have helped me with solving practical computer and robot problems as well as scientific dilemmas! Jørgen Bjørnstrup, Moritz Störring, Paolo Pirjanian, Claus S. Andersen, Bernhard M. Ege, Thomas D. Nielsen, Saber Sami, Vincenzo Marchese, Michele D'Agostino, Giuseppe Morana, and the entire group of the "maomao" friends; and then for their direct or indirect support at the beginning and during my scientific adventure in Denmark, Prof. Massimo Bergamasco, Prof. David Hogg, Prof. Roger Mohr, Dr. Jerome Blanc, Prof. Paolo Dario, Prof. Angelo Sabatini, Prof. Carlo Colombo, and Dr. Orazio Di Benedetto.

Finally, I would like to thank my family, (my parents, my sisters, my brother, Giovanni, my uncles Ausilio and Pio), and especially Kasia, for keeping the faith in me, and for having assisted and encouraged me.

This would not have been possible without you!!

Salvatore Livatino
Aalborg, May 2003

To my parents
for keeping the faith in me
and always supporting me with their love and care

Ai miei genitori
che hanno creduto in me
e che mi hanno sempre sostenuto con il loro amore

Contents

Chapter 1

Introduction: Robot Navigation, Problems, and Solution Models

Autonomous Robot Navigation is a research field which is very challenging and with enormous potentials. Despite more than two decades of research fundamental problems still remain unsolved, new technologies and proposed approaches push for new studies and exploration of the field, and nourish the hope that mobile robots could become part of our daily life by accomplishing a multitude of tasks (including services as well as entertainment), both at homes, offices, hospitals, but also industries, etc.

The purpose of this thesis is to investigate the field and propose a new method for autonomous robot navigation which benefits from the potential of the visual sensor to provide accuracy and reliability to the navigation process, while relying on naturally available environment features (natural landmarks). The goal is also to integrate techniques and algorithms (also related to other research fields) in the same navigation system, in order to improve localization performance and system autonomy.

The thesis content is described through 7 chapters aimed at scientifically presenting the main issues the thesis is dealing with, the proposed solution, argumentation, experiment, and learned lessons.

The thesis starts with introducing the proposed topic along with the related problematic of the field and commonly proposed solution models (chapter 1, Introduction: Robot Navigation, Problems, and Solution Models). The thesis then provides an overview of the most related state of the art (chapter 2, Mobile Robot Localization based on Vision and Natural Landmarks). The main characteristics and challenges of the proposed method are then presented and argued for, (chapter 3, The Proposed Approach).

The thesis proceeds with a more detailed presentation of the proposed approach to autonomous robot navigation through three chapters related to the primary system functionalities: Robot Self-Localization (chapter 4), Automatic Learning (chapter 5), and Automatic Recognition of Self-Learned Landmarks (chapter 6). In these three chapters the proposed solution models are addressed in their theoretical aspects, argumentation for their application to the proposed system is provided, as well as the results of performed experiments and related conclusions.

Finally, the summary and conclusions of the whole research activity are given together with some guidelines for future research, (chapter 7, Summary and Future Research).

Outline of Dissertation

The presented dissertation can in brief be outlined through the content of its chapters:

1. **Introduction: Robot Navigation, Problems and Solution Models**. This chapter provides an overview on the main issues in the field of Mobile Robotics focusing on autonomous robot navigation and localization. The main problems related to autonomous navigation are discussed, the most prominent "categories" of approaches and unsolved issues outlined, and an overview on the main state of the art solution models given. Eventually, the approach proposed in this thesis is introduced, which represents a solution to some of the unsolved issues.

2. **Mobile Robot Localization based on Vision and Natural Landmarks**. This chapter describes the most related state of the art through a review of selected representative contributions in the field of mobile robot localization based on vision and natural landmarks. The presented overview is intended to show the reader how the approaches proposed in the literature tackle some of the main issues in autonomous robot navigation through the use of sensor modalities and world representations which are related to what is proposed in this thesis.

3. **Proposed Approach**. This chapter presents the proposed approach to autonomous robot navigation by describing: proposed solution model, proposed choices, main challenges, and the implemented research development plan. A brief system demo is also provided thorough a "visual description" of the main system functionalities.

4. **Robot Self-Localization**. This chapter describes the first step of research suggested by the development plan, i.e Self-Localization. The proposed solution for robot-pose estimation is presented by the description of two fundamental steps: (1) automatic recognition of visual landmarks during navigation; (2) triangulation of recognized landmarks. The proposed self-localization scheme is then described, which integrates functionalities presented in both the two sections and add the Optimal Triplet Selection method. The performed experiments are then presented, and related conclusions summarized.

5. **Automatic Learning**. This chapter describes the second research-step in the development plan by presenting the main computational phases of the proposed learning strategy. In particular: (1) the acquisition of landmark candidate views; (2) the estimation of landmark position and orientation; (3) the analysis of the learned information in order to "refine" the learned environment model. The performed experiments are also presented. Eventually, the navigation strategy related to presented learning method is discussed.

6. **Automatic Recognition of Self-Learned Landmarks**. This chapter describes the third research-step in the development plan. It presents the proposed approach for automatic recognition of landmarks in case these have been learned by the method proposed in chapter 5. The chapter describes the proposed new method for automatic recognition based on techniques adopted from the field of Realistic Virtual-View Synthesis, which better exploit the possibilities offered by the proposed learning method. The performed experiments are then presented with the related conclusions.

7. **Summary and Future Research**. This chapter summarizes the main characteristics of the proposed solution model, method development, and main contributions of the thesis work. Open issues for future research are also discussed.

A. **Main Approaches in Realistic Virtual View Synthesis** (Appendix). The appendix A reviews relevant contributions on the field of Realistic Virtual-View Synthesis. It is intended for a reader interested in knowing more than what is described in chapter 6 about image- and model-based rendering approaches.

1.1 Mobile Robotics

In order to accomplish tasks a human being needs to posses the capability of understanding a problem and attempt its solution. In addition to this, if the human being possesses mobility and navigation capability, the range of performable tasks can immensely be extended. The same concept can be applied to robots[1], engineered systems which can be instructed by humans, in case these are required to autonomously operate in indoor dynamic environments.

As a consequence, mobility and navigation capability represents for a machine, and in particular for a robot, a huge potential, besides, very often, a real need. For example, a robot for cleaning floors needs to be able to clean as well as to navigate the environment.

The great potential and the need for mobility is the reason why the scientific research community has paid big attention to the study and development of mobile robots. In particular, a specific field called Mobile Robotics has represented a major research field in Robotics during the last two decades.

Robots are already a key component in the car industries to perform tasks such as: assembly, painting, welding, etc. However, an assembly robot is preprogrammed, does not move around, and the only thing it is good at, is to do the assembly. The main difference between this kind of robot and a robot which would be able to come into our houses is the motion capability, which in turn calls for autonomy and adaptability.

There are innumerable applications which one can think of for a mobile robot. In fact, mobile robots find application in all typical Robotics domains: assembly, packaging, painting, etc., and in addition, thanks to the mobility feature, possible application domains have also become: marine environments, air-space, hazardous or difficult access environments, and also indoor environments such as homes, offices, hospitals, etc., where a mobile robot can represent a "helping hand" to perform domestic, repetitive and tiring tasks, as well as a supportive assistant.

1.2 Navigation and Localization

Navigation for a robot can be defined as the capability of freely moving inside an environment being able to reach established target locations, called goals, while avoiding obstacles which may be encountered on the way to the goal. There are basically two approaches to mobile robot navigation: the behavior based and the model based.

Behavior based, reactive approaches involve little or no global planning and for example enable the robot to move along a corridor essentially balancing the optical flow on both sides of the robot, (Santos-Victor et al. [122], Coombs and Roberts [34]). The recognition of a particular behavior may provide information such as being approximately at the center of a corridor.

The model based approaches involve some level of geometric model of the environment, either built into the system in advance, acquired using sensory information during movement or a combination of both.

[1] The word *"robot"* was invented by the Czech novelist Karel Capek, [21], and later adopted by Isaak Asimov, [6], telling about the dream of building autonomous robots, willing, intelligent and human-like machines.

The model based navigation algorithms generally consist of three major elements:
(1) path planning, (2) obstacle avoidance, (3) localization. Path planning is the process of finding the way to the goal. Obstacle avoidance is the process of going around unexpected obstacles which can be encountered on the way to the goal. Localization is the process of finding the position and the orientation of the robot relative to an external coordinate system. The focus of this thesis is on localization.

The position and the orientation of the robot, in combination, are called the robot *pose*. Knowing the correct pose of the robot may represent a major need but it depends on the application. In some cases localization must precisely be quantified, (as in case of metric navigation), whereas in some other cases an approximate localization may suffice (as in the case of topological navigation). Some applications do not need any robot localization, though.

Consider for example a device such as an automatic swimming-pool cleaner or a lawn-mower, which both have been commercially produced. Either of these operate in a well defined region and needs to be able to identify the limits of the working region and avoid potential hazards while carrying out their own task. At the current state of design they do not posses the ability of computing their position during navigation neither of planning particular working paths. The behavior pattern is a random wandering, including the possibility to recognize some features like the pool edges or like being completely underwater or on a water edge. Theoretically, their performance can greatly be improved in terms of efficiency if these robots are able to compute their own position and thus plan a path to effectively cover the area. However, this possibility would represent an additional challenge which could not be necessarily justified. For example, these machines could run purely from the solar power and could then take the time they want to complete their job.

In other circumstances, it is instead important for a robot to posses knowledge of its pose. Many applications require in fact the robot to be aware of its current location during its ride to the goal, or to execute definite movements while interacting with other objects. Consider for example a factory or a warehouse, where robots are employed to fetch and carry, the localization knowledge can be useful to confirm that the robot is moving in the correct direction as well as to make the system able to avoid expected obstacles which can be encountered on the way to the goal. The localization knowledge is also useful to make the system able to position itself before interacting with another object. For example, a robot performing in a domestic kitchen needs to localize itself precisely to be able to grab the handle of the refrigerator door before opening it.

We can summarize that in the above mentioned different types of application, the robot needs to localize itself in order to operate. The position knowledge becomes then a necessary condition for successful completion of many of the tasks which a robot might be required to perform. In other words, localization plays a key role in various mobile robot applications.

In the literature, mobile robot localization has frequently been recognized as a key problem in robotics with significant practical importance. Cox [35] for example noted that "using sensory information to locate the robot in its environment is the most fundamental problem to providing a mobile robot with autonomous capabilities".

The robot pose can be expressed in relation to a global coordinate system, or relative to the robot system. In the first case localization is referred as global localization. Many approaches are incapable of localizing a robot globally; instead, they are designed to track the robot position by compensating small odometric errors. Thus, they differ from global localization, (the approach described in this thesis), in the sense that they require knowledge of the robot initial position; and they are not able to recover from global localization failures.

Probably, the most popular method for tracking a robot position is the Kalman filter, [76], [36], which represents uncertainty by single-modal distributions. While this approach might be very fast, it is unable to localize the robots under global uncertainty, a problem which Engelson called the "kidnapped robot problem", [50]. The main advantages of global approaches over local ones are: (1) the ability to recover from localization failures; (2) it is not necessary to specify the initial location of the robot. Global approaches consequently have the potential of providing an additional level of robustness and higher autonomy.

A combination of global localization and relative to the robot (local), can also be proposed such as, for example, the case when relative positioning is used to avoid collisions and global positioning is used to plan a working path.

1.2.1 Landmark based Localization

Typically during navigation, a robotic system computes its position relative to an *environment model*, i.e. a description of the robot workspace. This information can then be compared to the sensory input, so that any discrepancy can be used to correct errors inherent for example to the dead-reckoning system, (e.g., errors in wheel movement feedback due to slippage, etc).

Human beings use general landmarks, such as buildings and traffic lights to localize themselves during navigation. A similar strategy can be used by a robotic system. In particular, the use of landmarks for robot navigation and localization has become a popular alternative to having a geometrical model to the environment, through which to navigate and monitor localization. In this context, the environment model is a collection of landmarks which represent reliable references in the environment. Each match between the environment model and landmarks observed during the navigation, can be used to reduce the uncertainty associated with the robot pose.

There are numerous approaches to localization, there are different sensor modalities (e.g. vision, laser, sonar, etc.), and there are different application contexts (e.g. indoor, outdoor, industrial, domestic, office, etc.). The features used as landmarks can consequently be of different type.

The goal of robot localization can for example be achieved by using specialized landmarks or active beacons. They may be represented by: bar-codes, reflecting tapes, colored patterns, ultrasonic beacons, etc., which can be recognized by laser, vision and sonar system. Robots operating outdoor can make use of the Global Positioning System (GPS) to achieve the same goal, as long as the provided accuracy suffices for the application. Robots operating indoor and in environments not purposively structured for robot-localization may instead use other type of landmarks naturally occurring, such as natural objects: gateways, doors, windows, ceiling lights; to geometric features: corners, vertical edges, (for example arising by walls conjunctions); and to a combination of such objects.

The broad set of possible landmarks raises the feature selection problem: what feature produces the best localization result? The answer to this question mainly involves the issues of feature acquisition and their use for localization purposes. In other words, the localization problem consists of designing:

1. how features should be used to estimate robot pose? (i.e the localization problem);
2. how features should be learned and feed into the robotic system in order to maximize their exploitation and utility? (i.e. the learning problem).

This thesis addresses both the topics. A brief overview concerning different solution for landmarks proposed in the literature is provided in subsection 1.5. It is important to observe that if landmarks are defined as special structures already in the environment, then we have the possibility of autonomous navigation in unknown environments.

1.3 Autonomous Navigation

A main challenge in mobile robotics is autonomous navigation. This indicates the ability for a robot to perform navigation without human assistance. Further to this, if a robot is able to navigate in unknown surroundings, the system can be referred as fully autonomous.

But how much is there really a need for autonomous mobile robots in real-world applications?

Autonomy can be implemented to different degrees, where the navigation capability depends on the application. There is in fact a wide range of situations where autonomous capabilities are not truly essential and it is possible to aid the robot with human control, or to allow it to operate only under restricted circumstances. For example, a robot for transportation of golf clubs (e.g. the Wiz [132]) only needs to be able to follow a player who can easily be identified by a marker. The Mars rover (Stone [133]) as well as many other tele-operated robots did not "need" full autonomy either.

In other circumstances, despite it could still be possible to simplify the navigation task providing some external help such as a priori map or known space landmarks, it may be necessary for a robot to autonomously monitor its position in order to execute definite movements. For example, in a factory or a warehouse where robots are employed to fetch and carry objects, to place objects in shelves, to clean halls, etc., the most appropriate solution can be an autonomous and accurate navigation, (based for example on specialized landmarks or active beacons). Autonomous position monitoring may also concern inspections of high risk areas like a nuclear power plants. We can conclude that an autonomous monitoring of robot pose would make the navigation task much easier and more effective.

What kind of application does require a mobile robot which can navigate autonomously in unknown environments? (i.e. fully autonomous navigation)

Potentially, such a robot is required in many situations where direct assistance may be very little, tethered robots impossible (robot workspace is complex), tele-operation over wireless connections difficult, (TV signals to operator gets disturbed), etc. For example, while exploring remote or dangerous areas, despite the robot could make use of priori knowledge or external help, it is still necessary to enable the robot to navigate by itself to react to changes in the environment. This could be required during inspection of a dangerous part of a factory where an accident has occurred, or when a robot moves into unfamiliar parts of the territory. These time-consuming activities in unknown (but also changing) environments would be better performed by a fully autonomous robot than by a robot controlled by a remote human operator.

The general performance increase in terms of flexibility and adaptability provided by an autonomous robotic system able to navigate in unknown environments, would also be very beneficial in many other applications. Davison [39] hypothesizes the case of a "robot taxi" operating in a busy city. In this type of environment changes happen continuously, which lead to the fact that roads, traffic lights and other landmarks get modified. In case of humans in daily contact with the city, their internal map updates itself. In case of robots, we need a navigation system able to do the same.

In cases of hospitals, offices or homes, where mobile robots could be used to perform varied activities, such as post delivery, food distribution, cooking assistance, kitchen and bathroom cleaning, etc. but also activities for rehabilitation purposes or for support of elderly people like preparing meals, helping to get dressed, etc., the autonomous navigation capability becomes rather essential since the environment is highly dynamic. In fact, in dynamic environments objects change positions and people move around, which excludes the possibility of mapping the environment and totally rely on such a model. Furthermore, indoor environments can not make use of "naturally available" active beacons, such as represented by the GPS technology, neither it is convenient to use specialized landmarks. We can then observe, and summarize, that a major demand for the indoor robotic systems is autonomous navigation in unknown/changing environments, with very little human assistance.

In conclusion, a fully autonomous system, (i.e. a system able to autonomously navigate in unknown surroundings), can be defined as a system able to perform:

1. self-localization, i.e. the automatic estimation of the robot pose. In particular, the self-localization has the big potential to allow a robotic system to navigate without human intervention, thus providing the system with a high degree of autonomy.
2. automatic learning, i.e the automatic acquisition of the necessary information to perform the localization task. In particular, the automatic learning has the big potential to provide a system with high flexibility and adaptability to unknown/changing environments which greatly extends the application context.

The above statements extend the previous mentioned "localization" and "learning" problems (subsection 1.2.1) towards higher self-government. In particular, the above statements represent: (1) the autonomous localization problem; and (2) the autonomous learning problem.

1.4 Why is Navigation difficult?

The progress in the field of mobile robot navigation have been slower than it might have been expected from the excitement and relatively rapid advances of the early days of research. Certainly, undoubted progresses have been achieved throughout the years, as for example shown by the ability of the tour-guided robots RHINO [20], and its evolution MINERVA [140], which represent encouraging examples of mobile robots successfully accomplishing navigation of indoor environments not specially designed for them.

Nevertheless, the most successful projects still include a certain amount of human assistance, or have been proven in highly constrained applications. Systems where a robot is acting unconstrained in naturally occurring surroundings have often been proven only in very limited trials, or have not a simple application.

It is very interesting to understand the reasons why autonomous navigation in unknown surroundings is such a difficult task for robots, after all, it is something considered easy for humans or animals, who have no trouble moving through unfamiliar areas. The answer would seem to be that the complexity of the real world, even in the simplified form of an indoor environment, is such that robots have a hard time to achieve the required range of capabilities to cope with all situations.

Another aspect is the tendency of the designers of navigation algorithms to impose methods which use representations understandable by a human operator, but not always close to the sensor input. For example, the use of the Cartesian representation to denote the location of features and robot pose is usually not grounded to the way the sensors perceive the information. However, this often represents a convenient way to operate with the robot, to combine multi-sensory information as well as to understand the meaning of the involved information and make good use of it.

The physical construction of robots represents a further source of difficulty: problems with batteries, kinematic restrictions, power supplying, as well as the complexity of programming and communicating with the computer and all the devices on-board.

World Sensing

Model-based autonomous navigation in indoor environments requires a robot with the capability of monitoring its localization status, to confirm predictions and solve unforeseen situations. This means that we need to use information about the current state of the environment in order to navigate it.

It might be possible to get some kind of map of the house or office where the robot is operating to support navigation, however, it is unrealistic to think that everything can be known down to each detail. Further to this, indoor environments such as our houses and offices are dynamic, things change place and people move around. The environment consequently needs to continuously be observed and interacted, which in turn calls for external sensors. The world surrounding a robot needs to be sensed.

The sensor characteristics and its reliability is then a reason for autonomous navigation being a difficult task. All sensors are error prone and sensitive to noise, and no single sensor normally covers the whole range of operation. Main sensory systems involved in indoor navigation tasks suffer of different kind of problems, e.g. motion drift, unpredictable errors, low range coverage, sensitivity to certain materials, expensive computations, etc. The overall difficulty can be summarized as a reliable use of sensors and the combination of different sensor modalities.

World Representation

The world knowledge needs to be sensed, but importantly it also needs to be organized. This means an efficient environment representation allowing and supporting the navigation task. Unfortunately, world models proposed in the literature usually are too difficult and complicated to be generated, and they often only represent approximated models. This is mainly due to the complexity of the real world and the inaccuracy in acquired sensor information.

The popular choice for landmark based models, and in particular the use of *natural* landmarks, (i.e. special structures already in the environments), allows in principle for navigation of any indoor environment. However, there is not a definitive choice on which kind of natural landmark produces the best results for navigation.

Several examples have been proposed in the literature: proposed landmarks are different in type of contained features, in number of features for a single landmark, number of landmarks required for a safe navigation, etc. The choice for landmark has influence on the landmark acquisition and recognition strategy, as well as on the way robot pose should be computed. In this context the main difficulty is a reliable and automatic landmark acquisition and recognition.

Positional Information

The robot related positional information is usually estimated by algorithms which process sensory input and previously provided environment knowledge. Different algorithm have been proposed to estimate robot position and heading.

The main problem is to cope with uncertainties that necessarily arise when navigating in a complex and highly dynamic environments. The arising localization uncertainty is mainly due to noisy sensory inputs and approximated environment models. There have been many attempts in the literature to cope with these problems, for example by integrating multiple sensor readings.

Unfortunately, the arising errors are hard to model because of their non-systematic nature, which leads to the fact that the localization error lacks an effective error recovery method. In particular, inaccuracies allow the localization error to accumulate and grow in a way to prevent navigation for extended periods of time without a certain amount of human assistance. The main difficulty is also to model uncertain positional information.

Autonomous Behavior

Researchers have come up with a variety of successful modules to perform certain robotic navigation tasks. For instance, to safely round a known obstacle, to identify local hazard or to produce specific localization information. Joining these into complete systems has proven to be difficult, though.

It does not appear as if some kind of general overall algorithm will suffice to guide a robot in all situations, and evidence shows that this is not the case with biological navigation systems, (Ramachandran [115]), which operate as a collection of specialized behaviors and tricks. Of course humans and animals, who have been subject to the process of evolution while living in the real world, have necessarily developed all the tricks required to survive. This is not the case for robotic systems, yet!

The lack of reliable learning and exploitation of previously acquired information, in other words, the missing capability of learning the new and remembering the past, represents a major gap between biological and robotic systems, which prevents robotic systems from being fully autonomous. A considerable amount of human assistance is still required in state-of-the art systems and a fully autonomous behavior has only been proven for highly constrained applications.

In this context a main challenge is an accurate and reliable learning of environment model information, e.g. landmarks, and an effective use of this information during navigation. These challenges in turn represent the problem of reliably performing localization, despite the uncertainty in available environment model and robot positional information.

Summary

The main factors, responsible for making robot navigation, and in particular localization, a difficult and challenging task, can be summarized as: (1) complexity of the real world; (2) lack of a general approach; (3) need for human understandable representation; (4) physical robot constraints; (5) sensory system reliability; (6) model and landmark type; (7) error propagation control; (8) demand for autonomy.

This thesis will address problems related to points (5) through (8). These aspects seem in fact to be some of the main factors to be considered in the design of an autonomous navigation system. The points from (1) to (4) are instead considered outside the focus of this thesis.

In particular, the aspects analyzed in this thesis are:

- **World Sensing**. The world needs to be observed or, in other words, sensed. *(sensor issue)*
- **World Representation**. The sensed world knowledge needs to be organized and maintained. *(model and landmark issue)*
- **Positional Information**. The sensed world knowledge needs to be effectively exploited in order to provide localization information. *(positional estimation issue)*
- **Autonomous Behavior**. The robotic system needs to be self-reliant, in particular, able to learn the new and remember the past. *(autonomy and dealing with uncertainty issues)*

The issue of uncertainty is connected with all the above points. Hence, it is also a focal point to look at this issue while dealing with the above mentioned problems.

1.5 Approaches to Localization

How do we approach a solution to mobile robot localization?

During the last two decades of research in robotics, there have been many different methods and technologies proposed to tackle the problem of mobile robot localization. In general, the two main aspects to be addressed when designing a localization method are: extraction of *positional information*, i.e. the method for estimating robot pose, and *world representation*, i.e. the model of the workspace to which localization should be referred.

Both positional information and world representation involve the use of *world sensing*, i.e. the use of sensors to update the motion status of the robot, to observe the surrounding environment, etc.; and both these aspects are involved by the requirement of *autonomous behavior*, i.e. the capability for a robot of acting as independent and self-reliant. In summary, the robot localization involves: world-sensing, world-modeling, extraction of positional information and autonomous behavior.

In this section, the main characteristics of technologies and methods proposed in the literature for robot localization will be briefly reviewed following the breakdown of issues suggested in the previous section, (i.e. world sensing, world representation, positional information, and autonomous behavior). This section can be considered as the state-of-the-art ”answer” to these four issues, presented in the previous section as the main ”sources of difficulties”. Chapter 2 will then look more specifically into the literature concerning the use of vision and visual landmarks for mobile robot localization, since these represent the main characteristic of the approach this thesis is dealing with.

1.5.1 World Sensing: Sensors for Localization

There are different types of sensors proposed in the literature for mobile robots, from odometry to inertial, from compass to sonar, from laser to infrared, radar, visual, etc., (Jensfelt [72]). It is important to have a good understanding of the physical sensor characteristics because this will greatly increase the ability to exploit their potentials. In particular, it is important to have good models for the sensor since they are the input to the localization algorithms. In general, sensors for robotic systems can be divided in two groups: *internal* and *external* sensors.

Internal Sensors

The internal sensors are mainly represented by odometric systems, inertial sensors, and compass systems. These sensors measure some internal state of the robot such as: motion based on wheel rotations (odometry); acceleration, orientation, stationary conditions, (inertial); direction/heading (compass).

- **Odometry**

  Most of robotic mobile platforms are wheeled, so that the use of wheel encoders has become popular. In particular, knowing the size of the wheels and their configuration makes it possible to estimate the robot motion. The common term for this system is *odometry*.

Unfortunately, with this kind of system, errors accumulate so that even a small error may lead to a large error over time. An error may be due to wheel slippage or bumps. This kind of event may lead to a non-systematic error which can not be foreseen by the model.

In summary, odometry is a resource that should be utilized since it is known to be reliable over short distances (under normal circumstances), and it is usually available in most of the robotic platforms, (so that not using it would be a waste of information). However, care must be taken in order to not completely rely on it over medium (around 5 meters) to long distances.

- **Inertial**

  The are different types of sensors which fall in under the term inertial sensor. Three popular types are: accelerometers, gyroscopes and inclinometers.

  The accelerometers measure both linear and angular acceleration, they may also provide position information. Typically these sensors suffer a very poor signal to noise ratio for low accelerations. The gyroscopes measure angular and linear motion, they may also provide velocity and position. The inclinometers measure inclination gradient by electrically sensing gas bubbles moving inside a tube with liquid as effect of the gravity force.

  The inertial sensors can be used for dead-reckoning navigation and in robotics systems they can be complemented by an odometric system. This enables a reduction of the drift problem they are subject to. Inertial sensors represent a basic component of aircraft navigation systems, and they also find application in outdoor robots navigating in uneven terrains, (Oliveira and Santos, [111]).

- **Compass**

  The mechanic magnetic compass is the most common and oldest kind of magnetic compass. It uses the magnetic field of the earth to find the direction. The robot applications of compasses typically suffer from severe deviations caused by large quantities of metal, motors and high-power distribution systems, which are normally found on a robot. The compass is still a good source of information to estimate the initial robot orientation.

Different internal sensors can be combined to increase accuracy, reliability, and range of operation.

External Sensors

The external sensors are mainly represented by vision, sonar, laser, infrared, and radar. They provide information about objects in the workspace surrounding the robot, for example the distance to objects, or revealing the presence of a certain feature. In particular, external sensors may provide: shape and range information to short-medium distances (vision); the same also for longer distances (sonars), and with higher precision (laser); information about the presence of objects at short distances (infrared), or very long distances (radar); etc.

- **Vision**

  Vision is the sensor which has the greatest potential, but also the sensor which is most difficult to master. The visual observation is created by arrays of photo-sensitive cells recording light intensities related to a given workspace-portion.

  Even though vision has been studied for decades, few robust algorithms exist. The algorithms might perform very well under some conditions, but they probably will not work any more if light conditions or textures have changed. Most of the algorithms which perform well, are written ad hoc for a particular purpose and they lack generalization. Another problem is, that vision algorithms typically are computationally expensive, e.g. the case when vision is used to compute range information.

  The vision sensor is believed to be the most powerful sensor if its full potential is utilized. The vision great potential could be used to solve most problems in robotics, in particular navigation and localization. A camera can be installed in the working environment to observe a robot moving around, or it can sit on-board the robot. Different camera configurations have been proposed: monocular, stereo-couple, triplet, omni-directional etc., where cameras can be proposed fixed to the robot or sitting on a pan-tilt units. Recently, there is a growing interest to omni-directional cameras, (Boult et al. [17], Gaspar et al. [56], Paletta et al. [113]), and panoramic synthesis from multiple images, (Bunschoten and Krose [19], Livatino and Madsen [92]).

  Vision is the main sensory system this thesis proposes, because of the rich and discriminant information believed to be contained into captured images. The basic argument for such a choice is its potential for providing discriminant and accurate information to be used for robot localization.

- **Sonar**.

  The ultrasonic sensor, the sonar, uses acoustic energy for measurements which is created using a transducer. The ultrasonic transducer transmits a short ultrasonic pulse, which will be reflected by the environment objects surrounding the sensor. An ultrasonic receiver registers what comes back after a reflection. The received signal is commonly processed by measuring the time-of-flight, i.e. measure of the time from transmission to reception of the signal. The time-of-flight is depending on the speed of sound in air, and thus the temperature, humidity, air pressure, etc, may effect measurements. The knowledge of time-of-flight enables the computation of the distance to the target, which reflected the pulse.

  The main drawback with the sonar sensor is its wide beam of perception, which causes the fact that it is impossible from one reading to identify the object position within the beam. To resolve this problem several readings need to be integrated. For example, three sonar devices may be able to identify planar objects like walls as well as cylindrically shaped objects, (Sabatini [121], Di Benedetto and Livatino [10], Wijk et al. [147]). Some of the sonar drawbacks are: specular reflections, the possibility of a "crosstalk" when using multiple sensors or multiple robots.

  Nevertheless, the sonar is the sensor most commonly used in Robotics. Some of the reasons for its popularity, is that the sensor is widely available, inexpensive, and easy to control. Most of the commercially available robotic platforms are equipped with a ring of sonar sensors.

- **Laser**.

  The sensor called *Laser*, (standing for Light Amplification by Stimulated Emission of Radiation), used in robotic applications is the laser range finder. The laser is used to measure distances. The distance to an object can be achieved by measuring the time-of-flight or the phase shift, The distance is measured to a single direction, with high range and angular resolution. The high accuracy makes the laser a very useful device for robotic applications and it has promoted the development of laser scanning sensors. The scanning effect is achieved by mounting the laser on a rotating body or a rotating mirror, (Nashashibi et al. [108]).

  Compared to the sonar sensor the laser scanner is very expensive, and transparent to some materials such as glass.

- **Infrared**.

  The infrared sensor represents another group of sensors used to measure distance using light. In this case the emitted energy is infrared energy. The distance information is achieved by measuring the intensity which comes back to the receiver sensor.

  The energy is not as highly concentrated to a particular direction as in a laser system, giving a much lower range of coverage, typically less than 1m. This sensor is also very sensitive to reflecting materials. On the other hand this sensor technology is very cheap and robust for application where it is employed to provide occlusion informations, i.e. to detect whether there is something in the direction of the sensor or not.

- **Radar**.

  The radar sensor works under the physical principle of time-of-flight, phase shift and frequency modulation. It operates with frequencies around 3 to 300 GHz, which allow to combine long-range sensing and sufficient/high resolution. This is the reason which has made radar suitable for outdoor applications and very popular in air traffic control. The radar is also used in many other applications including automatic brake systems for cars, (Tullsson [146]), and outdoor robotics.

  The radar sensor may become sensitive to atmospheric attenuation, air pressure, temperature, humidity and surface reflectivity. The latter might cause an object to become invisible.

Different external sensors can be combined to increase over a single sensor accuracy, reliability, range of operation, etc. For example, there are some 3D laser imaging sensors available used in combination with sonars to exploring unknown worlds (Dudek et al. [47]), Leiva et al. [81]). The laser sensor in combination with vision may form a powerful sensor pair which gives both information about intensity and depth, (Jensfelt and Christensen [74], Arras and Tomatis [4], projects: Resolv [71], Camera [53], Michelangelo [86]).

Other examples show how laser transparency to some materials such as glass, become more robust when combined with vision (Ruggeri et al. [119]); the use of vision and inertial sensors for map building (Lobo et al. [95]); the use of sonar and infrared, (Hallmann and Siemiatkowska [63]), etc.

Summary

The table1.1 summarizes, for the described sensory systems, provided information, advantages and disadvantages.

SENSOR	INFORMATION	ADVANTAGES	DISADVANTAGES
Odometry	pose update	widely available, low-cost	subject to drift, unpredictable error (depends on trajectory & distance)
Inertial	acceleration, direction, inclination	practical	subject to drift and noise
Compass	direction	practical	subject to disturbs
Vision	various (pos., rec., etc.)	great potential, large information	complex, sensitive to illumination, imprecise, slow
Sonar	range	low-cost, practical	low angular resolution
Laser	range	high angular resolution	high-cost, sensitive to transparent materials
Infrared	range	robust to reveal occlusions	low-range coverage, sensitive to reflective materials
Radar	range	long-range sensing, good resolution	sensitive to atmospheric phenomena

Table 1.1: Main characteristics of the described sensory systems.

1.5.2 World Representation: Models and Landmarks

Models for Localization

The knowledge about the robot workspace can be described in many ways. Typically, it is represented by a model which is used to assist the robot during navigation. This case is usually referred to as *model based navigation*. The research activity has investigated different types of models, which may contain different degrees of details, varying from a complete CAD model of the environment to a simple graph of interconnections or interrelationships between the elements of the environment. The research activity has also investigated the possibility of not using a model at all. The latter is usually referred in the context of behavior based navigation.

Different paradigms for mapping indoor environments have been presented in the literature. These may be classified as: topological, metric, appearance.

- **Topological Maps**. The environment is represented by a connected graph, where the nodes correspond to places, like a room or a corridor, and the links correspond to connection between places, like a doorway, the stairs, or an elevator, (Meng and Kak [104], Santos-Victor et al. [123]). A topological map can also be organized as a hierarchical structure where for example at low level there are rooms, and at higher level the office building.

The advantage is that they are abstract and can be constructed without knowing the exact physical relationship between different nodes. These representations often also contain local metrical information for node recognition and navigation support. Various approaches may differ on: information associated with nodes, different types of nodes, compactness of representation.

- **Metric Maps**. The environment is in this case described by a set of elements which are spatially located within the environment. The position of each element is known in respect to the map. Typically, a map element is represented by a *geometric feature*, or by a *cell* representing a workspace portion.

  1. **geometric features**. These maps can have varying complexity depending on the application. The geometric features typically represent interior spaces by the geometric model of the environment and they can have different level of detail, (Tell and Carlsson [138], Ruggeri et al. [119]). A typical representation is a CAD model.
  2. **occupancy/cell grids**. This type of data structure, (introduced in 1985 by Moravec and Elfes, [107]), represents the robot workspace as a collection of cells which can be free or occupied. Cells may also incorporate a probability value, (a measure of the belief that the cell is occupied), or an uncertainty value (to include errors in the location of identified objects). The occupancy grid map is usually suitable for accumulating data from rangefinder sensors such as sonar and laser, (Ribo and Pinz [117]), and combine readings from cameras and rangefinder, (Chatila and Laumond [27]).

- **Appearance-based Maps**. The environment can be represented by raw sensor data. For example, a collection of images, laser scans, or sonar readings. This type of representation has the advantage that model elements are grounded in the perceptual abilities of the robot sensory system. Consequently, an appearance map usually represents a richer and discriminating information than the one represented in topological or metric maps.

  Since there is a large amount of data to be processed, only some salient subparts may be considered from the initial raw data, e.g. texture-paths, (Balkenius [9]). Special functions may also be applied to the raw input data, e.g. principal component analysis, (De Verdiere and Crowley [42], Winters and Santos-Victor [148], Zingaretti et al. [154]), grey-value invariants (Schmid and Mohr [124]). Many training images are usually needed to obtain an accurate model.

The above presented representation does not always represent alternatives. A combination of different paradigms is in fact possible. For example, Thrun [139] integrates a metric-based approach and a topological one. In particular, occupancy maps are first learned and then transferred into a topological representation.

The environment model has been considered as a fundamental support to many navigation tasks. However, the processing time needed to build up a detailed model and the error associated to model elements and sensor readings, still represent major problems. This has questioned the way a model should be learned and represented, and directed the research activity towards either simpler but still efficient world representations or *map-less* navigation schemes.

Landmarks for Localization

The environment can be represented by a collection of landmarks which usually have an associated "meaning", (e.g. they represent an object, a structure of the environment, etc.). In this case, the environment does not need to be entirely represented to each detail, as long as the amount of landmarks suffices for the purpose of robot localization. A landmark-based model is in principle very suitable for dynamic localization because it resembles the model humans use to navigate.

The number of landmarks in a map mainly depends on: sensory system, localization method, application context, and required performance. Landmarks may be of different types: artificial and natural.

- **Artificial Landmarks**.

  When a mobile robot operates in a static workspace, for example in an industrial setting, it is possible to engineer the environment in order to simplify the localization problem. This can be done by adding artificial landmarks to the environment. This solution can be very convenient since the cost for adding artificial landmarks is not significant in a setting where the environment will not change. The shape or pattern used by artificial landmarks makes simple and reliable to detect and identify them from sensor readings.

  1. **active beacons**. The active beacons systems can use radio signals, light sources, ultrasound technology, etc., to be identified. This allows for an easy recognition and landmark "visibility". Triangulation methods have been proposed to compute the absolute location of a robot in indoor environments, (e.g. based on infrared sensors), while for outdoor applications, a typical example of active beacon system used for robot localization is the Global Positioning System (GPS), which can be proposed in different versions with different position accuracy, [72].
  2. **passive beacons**. The passive beacons can be: circles with bar codes, (Kabuka and Arenas [75]), reflective tapes along a robot path, (Tsamura [144]), colored strips, special patterns, (Clarke [31]), etc. The passive beacons have the advantage to be a low cost solution while potentially providing an accurate and absolute position information. A disadvantage of passive beacons is that beacons need to be visible and recognizable, typically, by a vision system. However, also laser-based system, sonar and infrared, have been proposed.

- **Natural Landmarks**.

  When a robot is operating in offices or domestic environments it might not be acceptable to place artificial landmarks throughout the environment. In addition, installation of bar codes or reflective tapes typically requires a level of knowledge that an ordinary customer does not posses. Furthermore the demand for aesthetic in our homes is much higher than in a factory.

When using natural landmarks for localization, one of the problems is to find suitable candidates for such a role. The landmarks used by different approaches can be of different type. Among them: geometric features and object appearances.

1. **geometric features**. These type of landmarks are geometric features extracted from the environment surrounding the robot, e.g. vertical lines (which can represent a wall segment), corners (which can represent the intersection between walls), etc., (Neira et al. [109], Tell and Carlsson [138], Arras and Tomatis [4]). Those features are abundant in typical indoor environments and methods for their extraction from sensor readings are well known. A main field of investigation is then how to correlate geometric features in order to achieve reliable localization information.
2. **object appearances**. These types of landmarks represent objects typically occurring in an indoor environment such as doors, windows, door handles, light switcher, ceiling lamps, etc.; but also just part of them. For example, those represented in texture-patches, (Zingaretti and Carbonaro [152], Balkenius [9], Hayet et al. [68], Davison et al. [41], [40], Livatino and Madsen [94]). The idea is that a system could recognize the appearance of these objects and so use them for localization purposes. It is consequently important that such objects would be visible, viewpoint invariant, and discriminant.

Natural landmarks can be represented by a combination of different feature types. For example, the type of landmarks this thesis proposes (a visual landmark) represent structures of the environment which can be recognized in images from their appearance, nevertheless, the selection is based on a content of geometric features.

A landmark-based model does not represent a mapping method alternative to the previously presented ones (topological, metric, appearance). In fact, a landmark could represent a specific place of the environment and so it could be associated to a topological map, (Santos-Victor et al. [123]), or a landmark could be used together with a geometric model of the working environment. A landmark model can also be a combination of metric and appearance information. For example, the type of landmarks this thesis proposes, represent the environment appearance, but they also have associated a metric information, i.e. the landmark pose, (Livatino and Madsen [93]).

1.5.3 Positional Information: Robot Pose Estimation

As discussed in section 1.2, in the context of metric navigation the knowledge of robot pose is a condition for successful navigation and it needs to be continuously addressed during navigation. There are, however, different methods proposed in the literature for estimating robot pose, mainly depending on: the type of model representing the environment, the type of sensors in the robot, but also, application context and required accuracy.

The pose estimation approaches can be classified depending on whether the localization relies on the presence of an environment model or not. If no model information is available global localization is not possible and the robot's capability of navigating the environment can be considered based on a *map-less localization.* If on the other hand, model information is available (i.e. *map-based localization*), then the robot can localize itself either with some manual help (*manual localization*) or more interestingly in a more autonomous way (*automatic localization*). In case of map-based localization we have the possibility for a global or relative positioning. Different techniques have been proposed in the literature which are briefly summarized in the following.

- **Map-less Localization**. In this case no internal representation of the environment is available to the robot, so that the navigation is behavior based. Robot navigation consequently needs to rely completely on the information provided by its sensors. In particular, the sensory input can be interpreted to directly extract navigation and "positional" information, as in the case when a robot attempts to localize itself by directing its motion with respect to some other object in the environment or landmarks, (e.g. Balkenius [9]).

  Absolute positioning (or even relative) is not required so that this technique is more suitable for topological navigation. Among proposed navigation techniques, the prominent ones are vision based. In particular: optical flow and appearance. For example, in the work of Santos Victor et al. [122], the robot finds its location, (e.g. staying in the middle of a corridor), essentially by balancing the optical flow on both sides of the robot; in the work of Gaussier et al. [57], navigation is achieved by memorizing the environment visual appearance. This way the robot location is only seen in respect to the goal and images might be associated to commands and controls that will lead the robot to its final destination.

- **Map-based Localization**. In this case the model of the environment is instead available. The robot may rely on it to estimate its position. In particular the system could be manually localized or it can do the localization automatically.

  1. **manual localization / initialization**. Manual localization is typically performed with the help of the user and it is typically used to either initialize robot pose, or to compute robot pose in cases when the error accumulated during navigation is such that there is no other way the robot can recover this error autonomously. Depending on the specific context and required precision, robot initialization can be implemented in different ways.

     Note that the process of robot pose initialization can in some cases just be initiated by the user who then let the system estimate it. For example, in a previous work of the author, the technique proposed for robot pose initialization only required the user to make the robot "look at" a specific workspace region, (containing three green landmarks), to let the system calculating an accurate initial robot-pose, [93].

  2. **automatic localization**. The central idea in automatic localization is to provide a robot with some references or landmarks contained in the map, expected to be found during navigation. The main task of the localization algorithm is then to identify the references or landmarks, so that the robot can use the provided map to estimate its pose by matching observation and expectation. In particular, automatic localization can be performed by the robotic system by either updating a previous position estimate, (pose refinement), or by computing an absolute estimate anew, (global localization). An overview of major techniques proposed in the literature for pose refinement and global localization can be found in DeSouza and Kak [46].

- **pose refinement**. A popular example is the dead-reckoning system (odometry), which updates robot pose based on wheel rotation. More reliable systems presented in the literature use instead external sensors, for example to perform landmark tracking. In this case, the robot pose is updated by keeping track of a feature or landmark in consecutive sensor readings, (e.g. images) which are recorded as the robot moves. Typically, the localization algorithm must simply keep track of the uncertainties in robot position as it moves, (Crowley [36]), and when the uncertainties exceed a threshold, a new pose estimate is requested as well as a new "target" to track. The landmarks used can be either artificial or natural ones. In case of vision based systems, the visual tracking is carried out by using methods based on image correlation and landmark tracking, (Davison and Murray [41], Zingaretti and Carbonaro [152], Christensen at al. [30], Hashima et al. [67]).
- **global localization**. In global localization an absolute robot pose is required to be estimated by the system based on sensor readings. For example, by using artificial landmarks at known positions. An absolute estimate can also be obtained by combining multiple sensor readings, for example using sonars (Sabatini [121], Di Benedetto and Livatino [10]), or by observing multiple features or landmarks at known position, as in the case of a vision-based triangulation scheme, (Atiya and Hager [7], Betke and Gurvitis [12], Andersen and Gonçalves [3], Madsen and Andersen [97]), and as proposed in this thesis (see chapter 4).
  The triangulation has become a popular way of combining multiple sensor readings. In particular, the use of a combined sensor reading is very beneficial if this allows for an absolute estimate of robot pose (instead of a relative estimation). In fact, an absolute localization has the great advantage of recovering from global failures, (preventing errors to accumulate).

Chapter 2 will look more specifically into localization approaches related to the topics this thesis is dealing with.

1.5.4 Autonomous Behavior: Learn and Remember

Autonomy can be implemented to different degrees, where the navigation capability depends on the application. This section aims at briefly classifying robot navigation skills in terms of autonomous behavior, to better characterize the capabilities of the system proposed in this thesis.

The minimum degree of autonomy is represented by *pre-programmed navigation*, i.e. the robot performs established movements in order to accomplish its task. A higher degree of autonomy is *self-localization capability*, i.e. the robot is capable of "sensing" its surrounding environment to automatically localize itself. The capability of a full or semi -automated map acquisition for self-localization, then represents a further level of autonomy, named: *map learning capability for self-localization*. Eventually, the ability of *simultaneous learning and localization*, might represent the highest degree of autonomous navigation.

The possibility of navigation without any description of the environment, *i.e. map-less navigation* (DeSouza and Kak, [46]), is also briefly mentioned. Nevertheless, this thesis focuses on how to provide a robot with autonomous capabilities based on the use of the environment model. In other words, the focus is on the ability to *"learn"* a map and then *"remember"* it, for self-positioning purposes.

Pre-programmed Navigation

The simplest case is represented by a robot pre-programmed to perform specific movements in order to reach its goal. This case foresees little interaction with the environment, for example limited to a reaction to a hazard or an unexpected situation. The robot navigates with "closed eyes" and it might use internal sensors to update its current localization status.

Self-Localization Capability

A much wider degree of autonomy is provided by a robotic system able to compute its pose automatically, i.e. self-localization. The robot is provided with a map of the environment prior to its navigation, and it e.g. uses landmarks to determine positional information.

Robot pose is typically estimated based on external sensors. For example, these sensors may be used to identify landmarks which directly provide the robot with its position, (artificial landmarks), or they contribute to determine robot pose by matching observation (e.g. an image, current sonar readings, etc.), against the expectation (e.g. landmark description in a database, simulated sensor-readings, etc.). The environment model consequently needs to be transferred into the robotic system before the robot navigates the environment. This allows the system to posses a very complex and detailed map. However, this map can not be updated during navigation, so the robot must stop and load a new map if significant changes occur in the environment.

Map Learning Capability for Self-Localization

The robot computes its pose relative to an environment model. This implies that the robot needs to know the environment model before computing its pose. Since a great limitation to robot autonomy is to "externally" provide model descriptions, researchers have proposed robots capable of autonomously exploring their environments, and generating an internal representation of it. In particular, this exploration process can be automated (*automatic learning*), or semi-automated (*manual training*).

1. **manual training**. In this case an internal representation of the environment is created by the robot itself with some human assistance. In this way it is possible to avoid the workload of manually generating a map before the system is put in place to work, (for example, the construction of an appropriate CAD model). Manual training usually allows for building a map just before navigating the environment, (which means a higher probability of building a model based on current environment configuration), and using a world representation close to sensory input, (e.g. appearance-based).

   The operation of map-building can be performed by manually guiding the robot to learn its own landmarks, which has the advantage that landmarks can be learned by a convenient observation configuration. For example, in a previous work of the author

[93], the robot was guided along a pre-determined path in order to capture a frontal picture of landmarks, taken from a convenient distance.

2. **automatic learning**. In general, it is impossible to assume that a map of the environment can be provided, generated, or manually trained by a human, especially if one has to provide metric information. Relying on such information would thus significantly reduce the potential utility of the system. It would instead be very beneficial to make a robotic system able to use its sensors to automatically construct its own environment model, to be used for self-localization during navigation.

   An important issue is the way the automatic learning is performed, e.g. static or dynamic learning, known or unknown pose uncertainty, etc. There have been many researchers who have proposed automated or semi-automated robots that could explore their environment and build an internal representation of it, (Davison [41], Trahanias et al. [141], Zingaretti and Carbonaro [153]).

Simultaneous Learning and Localization

This represents the process of building a map at the same time as estimating the pose of the robot, typically denoted as *simultaneous localization and mapping (SLAM)*. Three main directions can be identified in the literature: topological, grid-based and feature-based approaches. The correlation between the estimate of the robot pose and the map that is being constructed usually needs to be explicitly modeled in these cases. Most of the existing approaches are based on Kalman filters. These usually performs well in small environments, but the computational complexity becomes too high for medium-large environments. There have been many researchers in the recent literature addressing a simultaneous localization and mapping, (Neira et al. [109], Jensfelt [73]).

Map-less Navigation

This case represents the systems that use no explicit representation about the space in which navigation takes place. In particular, no map is ever created. Rather, the system resorts to recognizing objects found in the environment, (e.g. walls, doorways), based on their appearance, or tracking those objects by generating motions based on visual observations. Despite a map-less navigation could represent the highest degree of autonomy for a navigation system, it has the limitation that the navigation can only be carried out with respect to the object recognized. Hence, these techniques do not allow for global localization.

1.6 What is still missing?

In the previous section titled Approaches to Localization, the main approaches related to the problem of mobile robot localization have been briefly reviewed. This section summarizes the reviewed approaches, pointing at desired characteristics which are still missing. This is done in order to provide the reader with a better understanding of: (1) important issues still to be solved; (2) the answer to the important issues proposed by this thesis.

World Sensing

The robot sensory system is required to "sense" the environment in order to monitor the robot localization status, for confirming predictions, and for solving unforeseen situations. Unfortunately, the need for accurate environment information and wide sensing areas, usually represent opposing demands when a fast and reliable result is required. There is not a single sensor which covers the whole range of operation and there is not a single or multi-modal sensory system which is reliable enough to guarantee autonomous navigation, (Jensfelt [72]). All the proposed sensory systems lack reliability in some circumstances and have operational limits.

Internal sensors work well only in the very short range of navigation because of the accumulated error. Consequently, they can not be trusted as the only source of information to compute robot position. In particular, odometric sensors are subject to the drift problem which leads to an unpredictable ever-growing error; inertial sensors are subject to drift and typically have poor signals to noise ratio for low range movements; and compass sensors suffer from severe deviations depending on robot metals and generated magnetic fields.

The above reasons call for external sensors as dominant modality for the navigation task. Unfortunately, external sensors lack reliability because of inaccuracies in sensor readings, often resulting in imprecise and unreliable localization responses. In particular, ultrasonic-sensors make it difficult to localize a sensed object within the wide beam and suffer from specular reflections; laser sensors have a high cost and they are transparent to some materials such as glasses; infrared sensors have a low range coverage and they are sensitive to the environment materials; and finally vision, the sensor modality believed to have the greatest potential, is the most difficult to master. Vision systems are sensitive to changes in illumination and only few robust algorithms exist, which typically are computationally expensive, or require a consistent amount of human assistance to operate.

The algorithms used for the analysis and the interpretation of external sensor information play an important role, so they are also claimed as responsible for unreliable localization. In conclusion, state of the art external sensors do not suffice for robot navigation, and they often need to be complemented with (assisted by) internal sensors.

What is missing is a dominant sensor modality which would allow for extracting a rich and reliable information, which would "drive" a high level the navigation process, allowing the robotic system to take decisions on where to go, how to proceed, etc. In other words, a sensor modality which supports the system in autonomously and continuously providing a reliable answer to the question: *where am I?*

World Representation

The environment model has been considered as fundamental support to many navigation tasks. However, the processing time needed to build up a detailed model, and the error associated with automatically learned models, still represent major problems.

Advances in 3D model reconstruction nowadays allow for a more practical and automatic generation of geometric models. For example, reconstructions based on range images, (see the Camera project [53], Resolv [71], and Michelangelo [86]), including surface completion of occluded regions, (Castellani, Livatino and Fisher [24]). Unfortunately, the processing time for model construction does not yet make the proposed techniques suitable for real-time applications.

A faster alternative represented by the use of topological models has only been seen as suitable for tasks like moving from place to place, i.e. not for a precise robot localization. For example, one of the major difficulties for topological approaches is the recognition of "nodes" previously visited.

As for grid models, which allow for an efficient merging of sensor information, they might become computationally expensive if an accurate metric localization is required. Appearance based models are in principle computational expensive, so that they typically need a reduction of "space dimension" to be efficient, such as for example when using principal component analysis techniques. However, this often represents a significant loss of information.

Landmark based models represent in principle a suitable model for dynamic localization (they resemble the model humans use to navigate). However, until now these models have only been seen on commercial products if they use artificial landmarks.

Artificial landmarks can only be applied to static or almost static environments, since they require re-structuring of the environment if a change occurs. The only type of artificial landmark which does not require re-structuring of a "terrestrial" environment is the successful GPS system. However, the GPS can not be used in indoor environments.

As for natural landmarks, they have a great potential since it is desirable to have a world representation composed by features already available in the environment. Unfortunately, the exploitation of natural environment features has not yet turned into a truly robust reference, and there is not a definitive choice on which kind of natural landmark provides the best result for robot navigation. The most promising way to improve landmark reliability seems to be combining different feature types in the same natural landmark. However, there is not an unanimous consensus on which landmark type, or combination of types, provides the best performance.

Finally, the requirement of a world representation which can be automatically learned by the robot and successfully used for self-localization, has not been fully satisfied.

In summary, what is missing is a simple but efficient world representation based on natural environment features, which can be self-learned by the robot sensory system, and then used during navigation to perform accurate and reliable self-positioning.

Pose Estimation

The importance and the need for robot localization in a wide range of mobile robot applications has been discussed in previous sections. Unfortunately, accurate and reliable robot localization is still hard to achieve.

Map-less localization is only suitable for specific robot behaviors, (e.g. run through a hallway or a door, and obstacle avoidance), and it appears not to be useful for precise robot positioning. For example, the kind of accuracy required to perform fine interaction with objects, such as grabbing an handle situated at a specific workspace location.

Accurate robot positioning is consequently mainly attempted through map-based approaches. In this case, it seems that it is only by a manual localization that a precise pose estimates can be obtained. Unfortunately, manual localization also is quite an ”unnatural” modality to estimate robot position, not suitable for autonomous navigation.

In the context of automatic localization, the technique named as ”pose refinement” can only be used for short distances. For example, the estimation of robot position information can not be totally based on internal sensors, such as odometry, since the error they accumulate does not make them reliable on the long range of navigation; and, as for the use of external sensors, for example, a comparison of consecutive readings related to the same environment feature, (such as performed by many landmark tracking techniques), the main issue is the capability of landmark recognition and reliable transition between tracked landmarks. In general, proposed systems require periodical positional ”re-sets”, achievable by either a manual localization or absolute pose computation.

Techniques for absolute pose computation could thus nicely be integrated into pose refinement. In this context the use of combined sensor readings has shown to be very beneficial for an absolute estimation of robot pose, (instead than a relative estimation). Unfortunately, accurate and reliable absolute positioning seems only to be achievable when using artificial landmarks. Absolute pose estimation is in fact very challenging when using naturally occurring objects or landmarks, as reliable reference. Even worse is the case where those landmarks have been automatically learned by the system. Global localization consequently represents a formidable problem, (nevertheless, if solved, it also represents an issue with a dramatic potential!).

Since, as seen in the literature, it is not possible to achieve a very precise positional information for a mobile robot, it would be desirable to have a system which, in addition to a fairly accurate robot pose, would also be able to reliably estimate the uncertainty associated to the computed pose. A great attention should then be paid to model the error propagation.

In summary, what is missing is an algorithm which reliably estimates an absolute robot pose and its associated uncertainty, preferably based on a self-learned environment model.

Autonomous Behavior

Despite the progress in the field of autonomous behaviors, mobile robots still have a very limited autonomy capability. Pre-programmed robots only perform well in static environments. Self-localizing robots fail on the short/medium range of navigation, (especially when they are based on natural landmarks), so that localization needs to continuously be re-initialized. Methods for model acquisition as well as methods for simultaneous learning and localization, still are impractical for realistic settings, and especially in this last case, (where the time constraint is more critical), the computational complexity is too high for medium or large sized environments. Finally, map-less navigation systems are incapable of providing precise positional information.

A main problem for the self-learning is extraction of environment characteristics which can be robustly recognized in the future, despite changes in environment condition or robot position. The learned characteristics also need to be tolerant to errors affecting robot position during the learning and the navigation phases.

Unfortunately, the ability for a robot to ”learn” natural features and then ”remember” them for localization purposes has only been proven at experimental level, which explains why mobile robots have not yet come into our houses or offices.

In summary, the main problem is the lack of autonomy on the medium/long range of navigation. In particular, what is missing is a method allowing for a practical and robust self-learning of an environment model and an accurate self-localization based on the self-learned information.

Conclusion

From the analysis of main problems related to world-sensing, world-representation, positional information, and autonomous behavior, it appears clear how these aspects are tightly connected. The main aspects involved in the design of a robotic system are linked to each other, so that it is not possible, for example, to propose a model to represent the world without considering how this model can be sensed, self learned and used to estimate positional information.

It is consequently strongly suggested, (as also seen in most work by other researchers), to approach a solution in this field by thinking about an entire system concept. A solution to the problem of autonomous robot localization should be addressed from a ”navigation *system*” point of view.

1.7 What is proposed? Combining Vision, Landmarks and Triangulation

This section described the core characteristics of a system for autonomous navigation. The system combines the use of vision, landmarks and triangulation. The proposed approach intends to combine aspects which have shown high potential, though some of them still lack reliability.

The proposed combination is applied to a robotic system equipped with a single camera that can freely pan 360° relative to the robot heading. Figure 1.1 shows the robotic system used for the experimentations.

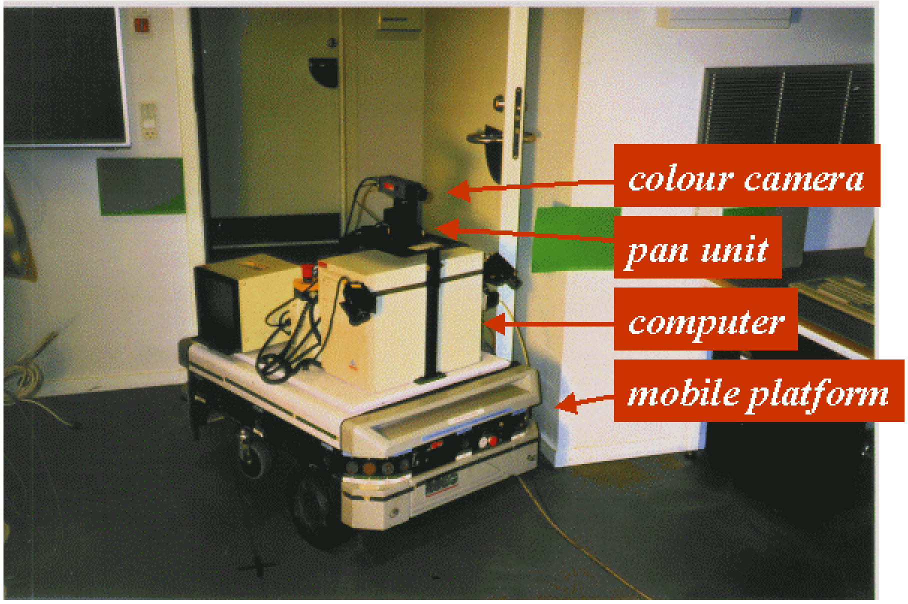


Figure 1.1: The robotic system used in this thesis.

During a learning phase, the vision system is used to learn representative views of natural landmarks and their position in the environment. During a localization phase, the vision system is used to recognize previously learned landmarks in the image plane. Every time three landmarks are identified in the camera image plane, the robot position and heading can be calculated by a triangulation method.

The proposed combination has the following immediate advantages:

- **Vision**. Possibility for sensing a wide area of the environment, and capturing discriminant information. The information pictured in an image is much richer than achievable with other types of sensors such as sonars. The visual appearance contains many more characteristics than sonar readings which only provide distance information. The use of vision is consequently suitable to represent discriminant characteristics of the environment (landmarks).

- **Landmarks**. Possibility for representing the environment by few but reliable representative environment features. The use of *natural* landmarks makes the system applicable to any environment without the need to structure it. In addition, the possibility to automatically learn and recognize natural landmarks makes the system fully autonomous.
- **Triangulation**. Possibility for reliably locating a robot on a medium-long range of navigation, and re-compute an absolute robot position and heading each time three landmarks are observed. The triangulation is a simple process and it does not involve any 3D reconstruction.

Beside the advantages, the proposed combination is also challenged by many drawbacks, which this thesis wants to address in order to eliminate some of them. The main drawbacks are:

- **Localization Error: Potential High Error**

  The computation of robot position and heading based on triangulation of landmarks can provide a very accurate estimate but it is shown to be very sensitive to noise, depending on spatial landmark configuration, relative position between robot and landmarks, and error associated with landmarks and robot poses. The effect of this noise is an error on the computed robot position, which may vary from few centimeters to many meters, (Madsen and Andersen, [97]).

  In particular, figure 1.2 shows how the noise sensitivity varies drastically with the choice of landmarks, (Livatino and Madsen, [93]). This clearly demonstrates the need for a strategy for selecting *good* landmark triplets, which in turn requires a technique for evaluating the possible landmark triplets before using them for triangulation.

- **Landmark Recognition**

  The landmarks to be used for robot localization can be any characteristic of the environment as long as they can be recognized. The recognition of landmarks previously acquired involves matching an observation to an environment model. This matching requires equal attention to both algorithm and representation.

  When using vision, landmarks may be represented by their visual appearance in the camera image plane. The use of a more robust algorithm, or a large landmark template, can imply a time consuming matching process. This choice could consequently slow down the pose updating rate so much that dynamic localization becomes impossible. On the other hand, a faster algorithm, or a smaller visual template, may cause ambiguous and misleading matches.

  An important issue is also represented by the way landmarks should be automatically learned and maintained in a model, and subsequently be processed and compared to current observation. In this case, the need for visual patterns which are invariant to observation viewpoint is in conflict with the desire of them being discriminant.

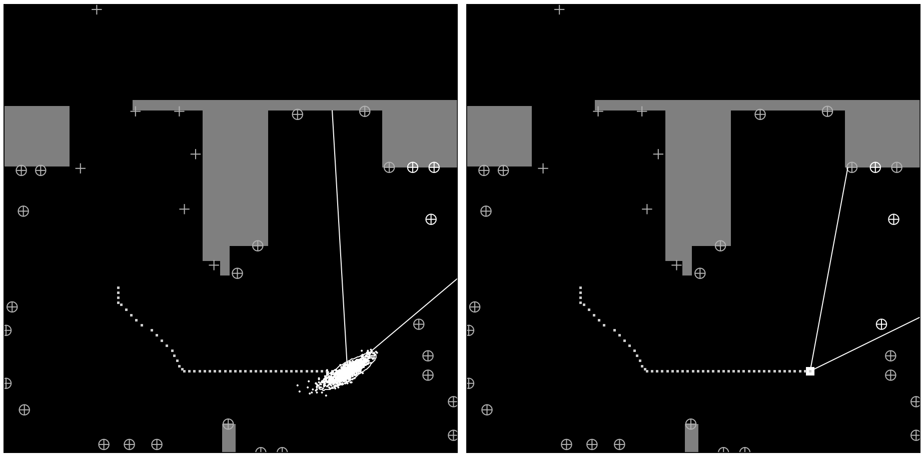


Figure 1.2: 2D floor map of robot workspace. Obstacles are shown in grey, free-space in black. The robot is following an assigned path, represented by small white points. All landmarks are shown as crosses. Visible landmarks are within a circle and those used for computing position have a white circle. The cloud of white points are robot positions computed with noise added to the sensory measurements. The two white lines emanating from the robot position indicate the camera field-of-view. The ellipse is the positional uncertainty computed by the 1000 simulated runs. The two figures represent variation in noise sensitivity when choosing different landmark triplets, but keeping the same robot position. In the right hand figure the chosen triplet provides a much better position estimate than that in the left figure.

- **Automatic Model Acquisition and Localization**

  The method for automatic learning and self-localization has to be designed specifically for the world representation we have chosen.

  The main problem is how to design the learning to guarantee that the acquired information can be successfully used during self-localization. In fact, environment conditions and robot position will likely be different during learning and localization, which challenges the landmark recognition, (different observation viewpoints, people may move around generating occlusions, ambiguous patterns may arise, illumination conditions may change, etc.).

  The learning of positional information related to acquired environment characteristics (required by the proposed method) will allow for a more flexible and re-usable world representation. However, the acquisition of positional information makes the learning process more complex and it raises the issue of how to obtain the positional information accurately and reliably.

  In addition, the error accumulated during robot navigation affects estimation of robot and landmark poses. Hence, it plays an important role in system localization performance. Consequently, the propagation of error estimates needs to be reliable.

The presented thesis work addresses the disadvantages related to the proposed combination of vision, landmark and triangulation, and describes a solution model coping with and/or reducing the system drawbacks, while exploiting the potential of the combination.

Chapter 2

Mobile Robot Localization based on Vision and Natural Landmarks

This chapter provides a brief overview on selected representative contributions in the field of mobile robot localization based on vision and natural landmarks. This overview is not to be considered as an exhaustive survey of the research field. Rather, it is intended to provide the reader with an overview of those methods which tackle the issue of autonomous robot navigation through a use of sensor modalities and world representations which are related to what is proposed in this thesis.

There are similarities between the described approaches and the one of this thesis. In fact, the works presented in this chapter have inspired the author of this thesis to undertake a work in this research field and to propose this type of approach and sensor modality. All described localization methods are based on vision and natural landmarks. Though, different camera technologies, types of landmark, and processing algorithms, are employed.

Each presented approach is described through a brief overview pointing out the proposed techniques and their main characteristics. A brief section at the end of each description compares the approach to the one proposed in this thesis.

2.1 Zingaretti and Carbonaro, [152], [153]

The method proposed by Zingaretti and Carbonaro relies on the use of vision, natural landmarks, and on the system learning capability for self-localization. The goal is to provide a robot with the capability of autonomously performing goal-directed navigation by means of a route following approach which allows for reliable detection and recognition of landmark features, [152], and including the capability of automatically learning the required landmark information, [153].

While a map of the environment is provided to the system prior to the actual navigation, the robot is asked to autonomously learn the type of landmarks required for its navigation. Landmarks consist of texture patches of 128 x 96 pixels representing planar surfaces already in the environment. In particular, landmarks represent orthogonal structures such as walls or ceiling portions. Landmarks are desired to represent the more discriminant environment features and the more invariant in a neighborhood. The characteristic of being discriminant is based on a *disparity map* obtained from stereo situation, (the robotic system is provided with a stereo camera head), by measuring angular discrepancy in position of an object in two stereo images. The invariant characteristic is established based on the *spatial activity*, (the sum of the difference of intensity information between adjacent pixels, edges, over a neighborhood).

The goal of the learning phase is the acquisition of landmarks which seem to be useful in some situations or in modeling the environment, and the selection of landmarks useful over the robot overall distribution of situations at certain positions. For this purpose five neighboring stereo images are selected in a certain configuration at chosen locations. The detected patterns represent visual landmark candidates and only a landmark matching procedure which runs during navigation will attest their usefulness. Landmarks should also be learned so as to allow the on-board cameras, observation of the next landmark.

The goal of the navigation phase is either to learn to ignore the target landmark or to scale and transform the landmark while keeping the same target landmark. In particular, during navigation the system makes up a list of preferred landmarks to approach, and then landmarks are selected based on current estimation of vehicle position. In order to improve tracking robustness the authors propose an *adaptive stereo template matching* where templates are generated by a genetic algorithm. The adaptive matching consists of adapting template processing parameters to generate optimal templates during tracking. A small set of landmark templates can then be used and landmarks can be detected in real-time. The fast computation allows for little changes to arise in target-object scale and point of view. The robot may alternatively be required to perform movements towards its target landmark for a better landmark observation.

Note that with proposed matching technique, instead of directly evaluating the raw intensity of stereo images, binary models from edge maps are used. The authors claim this yields to a better discrimination and noise robustness, and the system performance is robust towards illumination changes and reflexes in input images.

The result of landmark matching is the position of landmark center as well as its orientation. The computation of the orientation is simplified based on the fact that the only considered surface slopes are in the horizontal or vertical directions. The region of interest for performing the landmark template matching is calculated based on the 3D motion estimation. The motion is estimated by means of a feature based approach to motion analysis which matches template centers in stereo couples.

In summary, the main characteristics of the proposed method are:

- Landmark learning that deals with the problems of a good acquisition (discriminant and invariant patches), and/or efficient selection of potentially visible landmarks in following a route.
- Robot self-localization for goal-directed navigation based on tracking a sequence of relatively large landmarks.
- Landmark detection and tracking based on *adaptive stereo template matching*, based on binary models from edge maps and genetic algorithms, leading to discriminant landmark selection and fast processing.

Comparison

The proposed system has the advantage of exploiting naturally occurring features, no need for a precise estimation of robot and landmark position, the use of discriminant and invariant landmarks, and the capability of fast processing despite the relatively large landmark templates. Though, the binarization step prior to template matching is a critical operation which may result in a loss of information.

There are many similarities between the system described above and the one proposed in this thesis. Both the systems propose the use of vision and naturally occurring environment features (visual landmarks), represented by texture patches, and emphasize the characteristic for a landmark to be discriminant and invariant in a neighborhood.

However, the landmarks proposed in this thesis represent smaller texture patches and template matching is performed at appearance based level, rather than based on binary models of edge maps. Furthermore, the system proposed in this thesis does not rely on the tracking of one landmark, rather it exploits the recognition of three landmark features and relies on an accurate localization estimate.

The goal of the approach proposed by Zingaretti and Carbonaro is nevertheless represented by route following and thus differs from that of this thesis represented by accurate metric navigation based on global self-positioning.

2.2 Davison and Murray, [41]

The method proposed by Davison and Murray, [41], relies on the use of vision, natural landmarks, global localization and simultaneous localization and mapping. The goal is the capability of navigating in unknown surroundings by automatically building a sparse map of landmarks and relying on it to perform extended period of autonomous navigation in the same area.

While the robot is navigating the environment the system initializes features by fixating[1] them and running the operator proposed by Shi and Tomasi, [129]. This allows for the extraction of texture patches, (visual landmarks), sized 15x15 pixels each, representing high-contrasted textures. After fixation features are tracked, (one at the time), until when they remain in the camera field of view. At each computational step the system autonomously decides whether to initialize a new landmark or to keep tracking the current one.

While initializing, fixating and tracking landmarks, the system explicitly estimates the vehicle and feature positions relative to a world frame, and it maintains covariances of all the estimates. In particular, landmark position is computed based on odometry measurements, and by a stereo reconstruction technique (the system possesses a stereo head) which exploits epipolar relations and it is based on normalized cross-correlation. Variable search window (elliptical shape) are estimated based on current positional uncertainty and a Kalman filter based prediction scheme. The system is also able to re-find previously learned landmarks after a period of neglect.

The experimentation demonstrates the system to be able to autonomously perform localization while navigating. However, a stop may become necessary in case of sequential measurements of many landmarks, due to expensive correlation searches. Interestingly, the system estimates a sort of "observation area", that is, "a portion of the workspace where a landmark can be identified from every position". This workspace portion is defined by a range of allowed discrepancy in angle and distance between current robot position and the one at the time the feature was learned.

The main characteristics of the proposed method are:

- The use of a *single filter* to maintain estimates of all the quantities of interest (robot and features positions), and the covariance between them.
- Full uncertainty propagation (of robot and feature pose), where robot movements, feature fixations, shift-fixations, etc., are modeled as subject to Gaussian and non-Gaussian noise depending on the involved system components.
- Variable search windows based on positional uncertainty and "observation area" based on discrepancy in angle and distance between current, and previous (learning time), robot position.

[1] To *fixate* a feature is intended as the action of driving a camera to bring the feature to the image center.

Comparison

The proposed system has the advantage to alleviate the problem of motion-drift, exploit the potential of using active vision (tracking and fixating over extended period of time as well as wide field of view), transfer to visual domain of methodology used in directable sonar work (see [83]), absolute robot positioning (which can be re-set).

Similarly to the system proposed in this thesis Davison and Murray's work proposes a global localization method based on the use of vision and naturally occurring environment features (visual landmarks) represented by texture patches.

There are also many differences between the approaches. For example, no landmark tracking is proposed in this thesis, rather, a triangulation method based on three landmarks. The latter has the potential of providing high positional accuracy when combined with optimal selection of landmark triplets, however, the price to pay for the proposed triangulation is a more expensive computation due to at least three correlation-based searches.

Both the systems propose stereo matching of landmark template which exploits epipolar relationships and normalized cross-correlation. However, this thesis proposes larger template dimensions, a different operator to extract discriminant environment features, and a more complex estimation of landmark positional information, which includes landmark surface orientation, due to the estimation of landmark visual prediction. In fact, accurate pose estimates allow the system proposed in this thesis to estimate landmark *visual predictions* based on multiple reference views. The latter leads to "observation areas" expected larger than in Davison and Murray's work, (landmark image is not reprojected). However, in this thesis approach to determine landmark visual prediction is computational more expensive.

The work described in this section is characterized by the use of simultaneous mapping and localization approach (to maintain estimates of all the quantities of interest and the covariance between them), while the method proposed in this thesis propose two separate phases for landmark learning and self-localization.

2.3 Trahanias, Velissaris, and Garavelos, [141]

The method proposed by Trahanias et al, [141], relies on the use of vision, natural landmarks, and learning capability for self-localization. The goal is a method allowing for reliable navigation in unknown surroundings based on the system robustness to measurement inaccuracies. In particular, the method relies on accurate recognition of a previously learned sparse map of landmarks.

The approach proposed distinguishes between two phases: (a) extraction (learning) of naturally occurring environment features (visual landmarks); (b) recognition of landmarks during navigation.

During the extraction phase texture patterns are selected based on an attention selection mechanism which extracts salient image areas from "qualitatively segregated image-regions". Landmarks are then stored as raw patterns (appearances) along with associated informations (topological relations among landmark patterns). During the recognition phase texture patterns are selected by using an attention mechanism analogous to that proposed for landmark extraction. Hence, observed landmarks are matched to previously stored patterns. A viewing transformation is then proposed in order to improve matching performance. This transformation takes into account "observer" position during extraction and recognition time.

The proposed selective approach for landmark extraction and recognition starts dividing scene structures in partitions corresponding to walls, ceiling, floor, and far end. Then, edge detectors and adaptive Hough transformations are employed to determine the lines between walls and the floor. These lines will be detected in consecutive image frames in order to segregate image regions of interest on which to run proposed method for "focusing of attention", aimed at extracting distinctive patterns based on a linear combination of a chosen set of features. The choice of features depends on the application context. In case of workspace represented by orthogonal parallelepipeds (corridor-like) the authors propose the following features: area correlation over a window, image entropy over a window, standard deviation of the intensity histogram over a window, and standard deviation of pointwise differences across successive image frames. Visual landmarks are then extracted from 7x7 texture patches extended to a certain limit.

The proposed viewing transformation, denoted as *observer position transformation*, projects previously extracted landmark views according to current robot position. This allows for a more reliable image matching.

The experiments demonstrated the success of such a method on a corridor-like workspace, where the system is able to track corridor lower edges of the walls and confine the search for landmarks only to wall regions, (i.e. the environment portions believed more stable).

The main approach characteristics can be summarized as:

- Robot localization based on template matching of raw sensor data representing environment visual appearances.

- Extraction of distinctive image patterns based on an attention selection scheme including: qualitative space segregation, (edge detection, adaptive Hough transformations, dynamic segregation); saliency map, (linear combination of features), area of interest estimation, (large saliency areas, area expansion).

- Recognition based on a viewing transformation (*observer position transformation*) involving image transformation based on landmarks associated positional information and robot current position.

Comparison

The main advantages of the proposed method are represented by the capability of recovering from measurement inaccuracies by automatically selecting and recognizing distinctive image patterns. The benefit of a qualitative segregation including dynamic "edge following" seems, however, of immediate application only in corridor-like type of environment.

There are many similarities between the system proposed by Trahanias et al. and the system described in this thesis. Both propose a localization method based on the use of vision, natural visual landmarks, and a two phases navigation scheme, where texture patches are automatically learned by the system and later recognized during navigation.

Both the methods propose attentive schemes based on saliency maps calculated by combination of a set of features (though, through different computation schemes and feature sets), and they point out the need for a viewing transformation to enable matching of templates related to different observation viewpoints. In particular, the viewing transformation based on robot position both during learning and localization.

However, the system proposed in this thesis relies on accurate metric navigation rather than topological maps which rely on qualitative position information. In this thesis the available metric information is used to create (realistic) visual predictions, based on landmark estimated pose and multiple landmark views.

2.4 Balkenius, [9]

The method proposed by Balkenius, [9], relies on the use of vision, natural landmarks, and learning capability for self-localization. The goal is to provide a mobile robot with the capability of navigating in unknown environments by deriving "spatial representations grounded in the perceptual ability of the robot". In particular, the author proposes a hierarchy of spatial representations which enables the system to approach and recognize target locations, by means of "expectation-based" template matchings and "sensory-based" navigation behaviors.

The first representational level in the hierarchy contains landmark visual templates. Landmarks represent perceptual structures of the environment, (and not just physical objects), which contain a set of *features* (visual patterns) as distinct as possible, together with their spatial relations. The second representational level contains *view fields.* A view field represents a spatial structure. In particular, it defines a region of the space where a landmark can be identified from every position, (i.e. landmark visual template successfully matches). View fields are used by simple control strategies to make the robot approaching the *attractor* of the view field. This is the location towards which the robot should move to best recognize a landmark. Interestingly, movement direction can be derived from the perspective distortion observed in the template. The third representational levels in the hierarchy contains *place-fields.* These are combinations of view fields such as paths, etc. Finally, the forth level contains topological representations, to be used for planning, etc.

The proposed approach is implemented through two phases: exploration and navigation. During the exploration the robot constructs visual landmark templates for various locations of the environment. In particular, the system learns a set of distinctive texture patterns (feature) sized 7x7 pixels each, and their associated spatial relations.

During the navigation phase the system is required to recognize previously learned patterns, and the system uses the associated spatial relations to increase match reliability (feature and spatial relations have to be confirmed), as well as to speed up the processing (expectations on spatial relations constrain the search for the next feature and the process stops at the first "negative" response). Some discrepancy between observation and expectation is allowed, from this the name of *elastic template matching.*

The main approach characteristics can be summarized by the followings.

- Two phase navigation type of approach: exploration and navigation, where visual landmarks automatically learned during the exploration are recognized during the navigation.

- *Elastic* and *expectation-based* template matching, which allows for fast processing, and reliable recognition of visual templates at approximately correct locations.

- Navigation based on a hierarchy of spatial representations providing navigating behaviors from perceptual inputs.

Comparison

The main advantages are the capability of automatically recognizing distinctive templates, and by the characteristic of having perceptual grounded representations (appearances). The exploitation of the distance between selected visual patterns (features) in order to increase matching reliability and avoid expensive computations of large templates also represents an advantage. In particular, the use of spatial relations between features may compensate for the low discriminant characteristic of relatively small visual patterns in template features. Unfortunately, the proposed type of template requires an initial exhaustive search in the whole image for the first feature of a template, which represents a time consuming activity as well as a critical recognition step due to the small pattern dimension.

There are many similarities between the system described above and the one proposed in this thesis. Both the systems propose a localization method based on the use of vision, natural visual landmarks, and a two phases navigation scheme (learning and navigation).

There also are many differences between the approaches, including the type of visual pattern, world representation, and robot behavior. For example, Balkenius proposes an active approach to template matching, where the navigation process drives the robot towards a more reliable and "comparable" landmark observation. The observation is then matched to a prediction represented by a previously learned landmark template. In this thesis, it is instead the *visual prediction* which depends on current robot position (derived "passively" from previous learned template), and the current robot position can in principle be anywhere in the workspace.

Balkenius' concept of elastic template matching is in this thesis "replaced" by the concept of generating a realistic virtual-view as visual template "prediction". In this thesis the visual prediction is generated based on the knowledge of accurate positional information of robot and landmark, whereas in Balkenius' approach a "weaker" topological representation of the space is suggested.

The approach proposed by Balkenius seems consequently suitable for a behavior based navigation, whereas the one of this thesis is more oriented towards an accurate metrical navigation, (allowing for example precise interaction with environment objects).

2.5 Neira, Ribeiro, and Tardos, [109]

The method proposed by Neira et al., [109], relies on the use of vision, natural landmarks, and simultaneous learning and localization. The goal is to provide a mathematical framework for maintaining correlated features involved in the process of autonomous robot localization. The correlated features are in this case represented by environment features and robot location.

The authors investigate the limitations of using a priori maps (classified as incomplete, incorrect and imprecise). Hence, they underline the advantage of making the system able to automatically learn an environment map through its sensors during navigation. Being the estimation of both learned map features and robot locations, based on sensory information subject to errors, there is a need for a mathematical tool, (usually referred as *stochastic map*), allowing for modeling the uncertainty in map features and robot location.

The authors point out the issue of maintaining ***statistically correlated*** features (the feature estimates as well as the robot location, are in fact based on the same sensory source), and they propose a solution prototype. The proposed solution is represented by the *SPmodel*, a probabilistic model to represent and integrate uncertain geometric information that is specially suited to build and update stochastic maps. In particular, the SPmodel includes the estimates of the locations and the associated error.

The 2D version of the SPmodel and its application to stochastic map building is described. The proposed implementation concerns with a 2D monocular vision system used to determine the identity of the vertical edges observed in an image with respect to previous observation as well as with respect to the built map. In particular, extracted vertical edges are used to establish a correspondence between their associated projection rays and the features that constitute a map, i.e. wall corners and door frames. A matching process is proposed based on: (1) global coherence, i.e. that decide the matching for all the projection rays in an image; (2) feature tracking, i.e. vertical edges are tracked in the sequence of images that the robot takes as it moves.

The location of environment features and robot are estimated from a set of partial and imprecise observations, by including a version of the extended Kalman filter in the SPmodel (the estimation of feature location is nonlinear due to the existence of orientation terms). Tracking is used to estimate the motion of vertical edges in the image sequence. Edge prediction between adjacent image frames is in this case based on a Kalman filter. The Mahalanobis distance metric is used to determine whether two measurements can be considered compatible.

The main approach characteristics can be summarized by the followings.

- The use of the *SPmodel*, as a mathematical tool to represent and integrate uncertain geometric information (arising from uncertain sensory information).

- Feature correspondence problem based on global coherence based on the use of an extended Kalman filter for estimating prediction of feature location.

- Feature tracking based on a Kalman filter for estimating match between correspondent vertical edges in consecutive image frames, (simultaneous localization and mapping).

Comparison

The main advantages of the proposed method are represented by the capability of alleviating the robot motion-drift problem by building a simple vertical-edge based map. The feature tracking reduces the complexity of the correspondence problem allowing for fast computation, thus close views acquisition. An advantageous aspect of the proposed probabilistic model is that it can be applied to the case of a map built using a combination of sensors such as for example stereo vision and laser. Location of vertical edges should nevertheless always be within the range covered by the camera, (vertical edges always visible), which is critical during robot turns or fast motion.

The similarities between the system described above and the one proposed in this thesis mainly consist on the use of a monocular vision system and natural landmarks. On the other hand, there also are many differences between the approaches. For example, this thesis proposes a two phase autonomous robot localization approach, whereas Neira et al. propose a simultaneous learning and localization. In addition, in Neira et al. landmarks are represented by geometric image features such as the edges at a wall corners or door frames, whereas in this thesis landmarks are represented by distinctive texture patches.

The environment map in the Neira et al. work is simpler than the one proposed in this thesis since it is only based on vertical edges. This leads to the fact that the quality of the map is low for purposes other than correction of angular errors. In addition, vertical edges represent very ambiguous patterns, so that the system needs to track landmarks in consecutive image-frames and calculate edge-predictions in order to disambiguate them. On the contrary, in the approach proposed in this thesis landmarks represent very distinctive environment features so that there is no need for tracking.

From the computational point of view, the use of edges as landmarks and feature tracking allows for a fast matching, whereas in this thesis a more time-consuming matching process is required. The same can be observed for what concern landmark acquisition, being the vertical-edge detection faster than the proposed attention selection mechanism.

The two approaches share the "belief" that vertical edges represent useful features for robot localization. These edges may in fact well constrain camera rotation, hence, they may help to correct robot orientation. In particular, in the work of Neira et al., the detection of vertical-edges is used to correct angular estimates, and angular estimates are the main source of errors in the odometric system. In the approach proposed in this thesis vertical edges indirectly play the same important role, since they represent one of the feature considered by the proposed attention selection mechanism, and they are among the ones which are weighted most.

2.6 Paletta, Fintrop, and Hertzberg, [113]

The method proposed by Paletta et al., [113], relies on the use of vision, natural occurring "landmarks" (*visual context*), and self-localization capability. The goal is a robot localization method which is robust to occlusions and to noisy or uncertain environment information, based on the analysis of the visual context. In particular, a method is proposed to take advantage of the *local context* in omni-directional sensing by reasoning on the structure of ordered "unidirectional" views in the panoramic image.

The system is first trained through panoramic omni-directional images (size 360 x 200 pixels each), learned from different environment positions. Panoramas are then partitioned into *appearance sectors*, i.e. a fixed number of overlapping unidirectional camera views, (sized 20 x 200 pixels each). Then, after an image normalization (to account for illumination variations), the technique of Principal Component Analysis is applied to the images of a corresponding environment area to determine the most prominent eigenvectors. Each sector image is then projected to a multi-dimensional eigenspace, so that groups of sectors belonging to the same panorama form a *trajectory*, (i.e. the curve passing through the points representing appearance sectors). Distribution of the sectors in the eigenspace are represented by a mixture of Gaussian to provide a distribution over potential locations.

Recognition is supported by the property that close points in the eigenspace correspond to similar appearances. The recognition process is based on: (1) finding evidence for a position from sub-windows of the original image; (2) resolving the ambiguity in local sector interpretation as well as occlusion defects, by "Bayesian reasoning over the spatial context of the current position".

An interesting aspect is the way the system deals with the problem of partial occlusions. Occlusions may in fact induce ambiguity in single appearance sectors. However, occlusions hardly represent sector trajectories (in eigenspace) comparable with the training structures. The use of proposed method allows then for classifying panoramic input using "local window" data (the sector image), so that the recognition may remain optimal even in presence of severe occlusions (up to 70% occlusion in case of presented experimentation).

The main approach characteristics can be summarized by the followings.

- A sensor based model of the environment, (image based navigation), and eigenspace analysis applied to panoramic sub-images, (*appearance sectors*).
- The exploitation of *visual context* in omni-directional panoramic images to understand about the spatial relation between unidirectional views in panoramic images.
- The proposed Bayesian framework to enable to quantify the uncertainty in the discrimination of learned robot positions.

Comparison

The main advantages of the proposed method are represented by the capability of accurate visual navigation of autonomous robots in presence of severe occlusions, such as occurring in crowded places (offices, factories, urban environments), based on the characteristic to take advantage of even partial evidence about a position and to reject false hypothesis. In addition, the capability of implicitly recovering the robot's orientation by localization. However, the accuracy of estimated localization information is depending on the spatial density of trained sample, leading either to a complex system training or lower spatial accuracy in estimated position.

The similarities between the system described above and the one proposed in this thesis mainly consist on the use of panoramic views, and natural occurring visual appearances as "landmarks" for robot localization. There are nevertheless many differences. Among them, the spatial resolution of the considered panoramas, hence, of the "visual landmarks", and the recognition method.

The localization method based on Principal Component Analysis enables for significant reduction of processing and interpretation time. However, the accuracy of robot localization depends on the spatial density of trained sample, whereas in the method proposed in this thesis, it is relying on robot and landmark positional accuracy.

The relatively large portion of the environment represented in an image sector is challenged by occlusions which are likely to arise, so that this becomes a major issue, whereas an occlusion of a landmark in the method proposed in this thesis, may easily be overcome by using a different landmark. On the other hand, image sector recognition is in general more invariant to changes in viewpoint than in case of medium-sized landmarks (as those proposed in this thesis), since the different space resolution.

The goal of the approach proposed in Paletta et al. is nevertheless represented by topological navigation and so differs from the one in this thesis represented by accurate metric navigation. In the latter, a higher image resolution is believed necessary.

2.7 Yuen and MacDonald, [150]

The method proposed by Yuen and MacDonald, [150], relies on the use of vision, natural landmarks, and self-localization capability. The goal is to detected environment features automatically, and based on this knowledge to reveal the presence of obstacles as well as to correct robot localization errors.

The authors provide a general description of a panoramic stereo imaging system with only one optical center, and an image analysis procedure which detects obstacles by computing the distance to them. A number of implementation specific problems are then discussed and experimented.

The proposed system relies on *single optical-center panoramic views*, synthesized by cylindrical projections and matched by stereo-triangulation. The stereo situation arises from robot movements during navigation. The environment features extracted by the system are corners and edges. They are extracted using a Canny operator. The extracted features are used to identify image-regions of interest (around the features), which are then matched in correspondent stereo views. The match of the areas of interest is searched along cylindrical epipolar lines, (the camera position is known). The searching algorithm has a variable window based on the Shirai method, [80]. The match is based on the *minimum dissimilarity measure* between the template and the search window. The acquired images (synthesized in panoramas) are sub-sequentially matched in order to refine initial estimates by incremental updates.

The concept of *principal viewing axis*, i.e. the axis which is ”the extension line from one camera optical center to another”, (the two optical centers in a stereo configuration represent consecutive positions reached by the robot), is used as measure of ”trust” to improve from poor observations. In particular, knowing that the disparity estimation becomes more unreliable when the object being observed is getting closer to the principal viewing axis, an *incremental* update technique based on observations obtained in more ”favorable” conditions, is proposed.

Other than the incremental update based on principal viewing axis, the authors propose a *rotation correction algorithm*, based on the measure of angular deviation between specific image regions expected to be identical. These regions are in fact chosen near the principal viewing direction.

The experiments show panoramic views of 900x90 pixels representing 360 degrees views of the environment, acquired from known camera positions. The robot is let to run along short paths (few meters) plenty of rotations. The proposed method is demonstrated able to recover from odometric errors in estimated robot orientation (arising after rotations). The experiments also show a decreasing trend in 2D positioning errors based on the proposed incremental updating rule.

The main characteristics of the proposed method are:

- The use of *single optical-center panoramic images*, which takes advantage of the naturally occurring image sequences and the arising stereo situation with a baseline larger than conventional stereo heads.
- The development of an *incremental updating rule* to gradually replace "unfavorable" estimates.
- A *rotation correction algorithm* based on the redundant information associated with the principal viewing axis.

Comparison

The advantage of Yuen and MacDonald's techniques is to enable a reduction of the large uncertainty associated with vision-based positional estimates. Interestingly, the authors propose exploitation of the information associated with the principal viewing axis, and the expected similarity of the image regions near the principal viewing direction (e.g. around the *epipoles*). However, the proposed technique is not shown to allow for accurate measurements, i.e. it seems more suitable to detect obstacles rather than navigation landmarks.

The described method is similar to the one proposed in this thesis for what concerns the proposed vision system. In particular, the single optical-center panoramic view synthesis, the exploitation of the epipolar constraints, and the stereo triangulation arising from subsequent images acquired during environment "exploration". All those aspects in fact confirm the advantage of the system proposed in this thesis, and they show that the system proposed in this thesis is an up-to-date approach (the Yuen and MacDonald contribution is very recent).

The authors, nevertheless, address the implementation of specific issues, while the approach proposed in this thesis deals with an entire system concept, which includes automatic learning, recognition, and localization. Furthermore, the Yuen and MacDonald's approach seems only to be suitable (at this stage) for obstacle detection, while the method proposed in this thesis aims to learn accurate landmark positional information for robot self-localization.

2.8 Summary and Analysis

Summary

The described approaches present potential solutions to the problem of autonomous robot navigation, and in particular robot localization. The selected works often represent popular choices concerning type of landmark (e.g. geometric features), proposed technologies (e.g. omni-directional cameras), etc., so that they should not be considered as unique contributions in the field of that type, but more as typical examples of a certain approach.

The presented approaches propose different solution prototypes. Some of the authors propose a two-phase localization (Zingaretti and Carbonaro [152], Trahanias et al. [141], Balkenius [9], Paletta et al. [113]); while others propose a simultaneous localization and mapping scheme (Davison and Murray [41], Neira et al. [109]). There also are different types of landmarks proposed. Among them: geometric features (Neira et al. [109]); texture patches (Zingaretti and Carbonaro [153], Davison and Murray [41], Balkenius [9], Trahanias et al. [141]); visual context (Paletta et al. [113]).

There have been described different types of camera technology and processing applied to sensory information for prediction and recognition of landmarks. For example, concerning cameras: monocular systems, stereo heads, omni-directional sensors; concerning the prediction techniques: projection of image textures, adaptive matching parameters, etc.; and concerning the landmark recognition, matching techniques have been based on: image correlation, comparison of image features, etc.

The level of positional accuracy in robot and landmark locations, required and estimated by the systems, also differs. Different may also be the way the workspace is proposed to be modeled (topological, metrical, appearance-based).

Furthermore, the focus of the authors' contributions have often been on different aspects involving the localization issue. For example, Neira et al. [109] focus on modeling the uncertainty in case of statistically correlated features; Trahanias et al. [141] focus on the attention selection mechanism and recognition based on viewing transformation. Other authors focus more on the type of landmark template and spatial representation, (Balkenius [9]); matching techniques for discriminant and invariant patterns, (Zingaretti and Carbonaro [153]); quantifying the uncertainty in the discrimination of learned robot positions based on the visual context, (Paletta et al. [113]); exploitation of monocular system and panoramic views for error correction, (Yuen and MacDonald [150]); etc. The table 2.1 summarizes the main characteristics of the literature approaches revised in this chapter.

Analysis

All the presented approaches share the same common objective: a method which allows for robot navigation over extended period of time. In particular, the presented approaches propose the same technology and "navigation concept" as solution: vision as dominant sensor modality, and naturally occurring environment features as navigation landmarks. It is then quite clear that combining vision and natural landmarks is believed very potential.

On the other hand, it is not straightforward which is the most suitable technique one should adopt to estimate robot localization. There is, in the revised works as well as in others from the literature, a clear attempt of exploiting spatial relations, architectural structures, robot Kinematic, and other type of information, to validate the available sensor readings, and so improve localization estimates. For example, landmark tracking during robot navigation is often proposed to refine localization estimates, (though, the tracking can not provide alone global localization information).

A system based on the use of vision and natural landmark, which proposes accurate global localization with the purpose of autonomous robot navigation, can however hardly be found in the literature. In particular, it is difficult to find such a system described in all its main components (e.g. acquisition, map-learning, prediction, matching, localization, error propagation control, etc.). Lots of the literature works in fact mainly address specific issues.

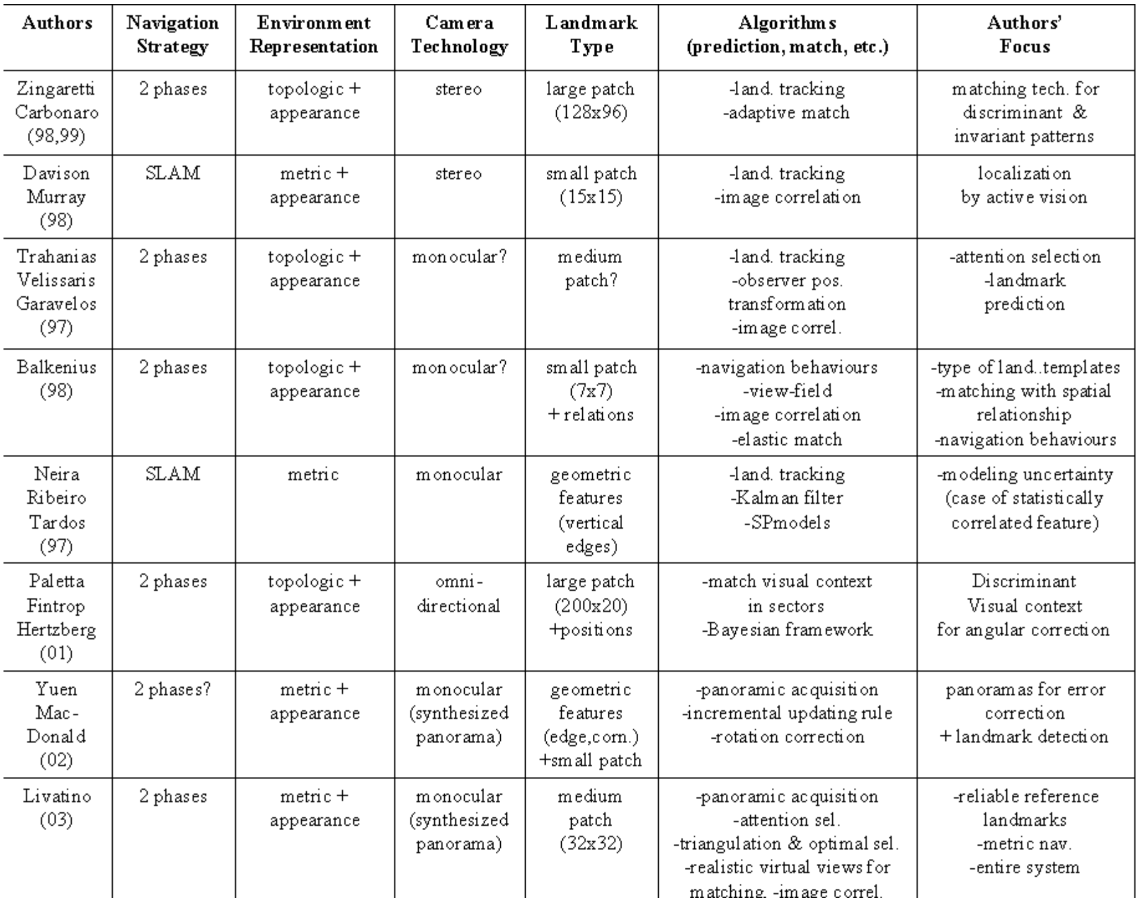
Though, the issue of proposing an entire autonomous system concept is also a very important aspect. In fact, from one side to propose a solution for a specific issue is very interesting since this can be applied as part of more complete systems, but from another side this confirms that to address a method for the entire system is difficult and complex.

This thesis consequently aims to propose:

1. an approach to mobile robot navigation based on the combination of vision and natural landmarks, which allows for accurate global localization;
2. an approach description for the entire (navigation) system main components which allows for autonomous navigation.

The proposed localization method is based on the robot-centered triangulation technique. The potentiality of the triangulation solution for robot localization has been demonstrated in previous literature works, (Betke and Gurvitis [12], Atiya and Hager [7], Andersen and Gonçalves [3], Feng et al. [52], Madsen and Andersen [97], Greiner and Isukapalli [62]), which have, however, often provided only the theoretical foundations and not a full system description for autonomous robot navigation such as this thesis aims to do. The table 2.1 bottom-row summarizes the main characteristics of the approach proposed in this thesis (according to the presented literature works "summaries").

The various contributions described in this chapter were intended to provide the reader with an overview of most prominent research approaches in a field related to this thesis. This in order to provide the reader with better understanding (and appreciation), about differences, potentials, advantages, drawbacks, and innovative aspects, this thesis proposes, which are described in the next four chapters.



Authors	Navigation Strategy	Environment Representation	Camera Technology	Landmark Type	Algorithms (prediction, match, etc.)	Authors' Focus
Zingaretti Carbonaro (98,99)	2 phases	topologic + appearance	stereo	large patch (128x96)	-land. tracking -adaptive match	matching tech. for discriminant & invariant patterns
Davison Murray (98)	SLAM	metric + appearance	stereo	small patch (15x15)	-land. tracking -image correlation	localization by active vision
Trahanias Velissaris Garavelos (97)	2 phases	topologic + appearance	monocular?	medium patch?	-land. tracking -observer pos. transformation -image correl.	-attention selection -landmark prediction
Balkenius (98)	2 phases	topologic + appearance	monocular?	small patch (7x7) + relations	-navigation behaviours -view-field -image correlation -elastic match	-type of land..templates -matching with spatial relationship -navigation behaviours
Neira Ribeiro Tardos (97)	SLAM	metric	monocular	geometric features (vertical edges)	-land. tracking -Kalman filter -SPmodels	-modeling uncertainty (case of statistically correlated feature)
Paletta Fintrop Hertzberg (01)	2 phases	topologic + appearance	omni-directional	large patch (200x20) +positions	-match visual context in sectors -Bayesian framework	Discriminant Visual context for angular correction
Yuen Mac-Donald (02)	2 phases?	metric + appearance	monocular (synthesized panorama)	geometric features (edge,corn.) +small patch	-panoramic acquisition -incremental updating rule -rotation correction	panoramas for error correction + landmark detection
Livatino (03)	2 phases	metric + appearance	monocular (synthesized panorama)	medium patch (32x32)	-panoramic acquisition -attention sel. -triangulation & optimal sel. -realistic virtual views for matching, -image correl.	-reliable reference landmarks -metric nav. -entire system

Figure 2.1: The table summarizes the main characteristics of the revised literature approaches. The bottom row refers to the approach proposed in this thesis.

Chapter 3

The Proposed Approach

This chapter presents the proposed approach to autonomous robot localization. First section describes the proposed solution model by addressing the proposed choices for what concern the main factors to be considered in the design of an autonomous navigation system, (see section 1.4). The second section will provide a brief system demo thorough a "visual description" of main system functionalities. The third section will focus on the approach main challenges. The proposed research development plan is described in the last section.

3.1 Proposed Solution Model: System Concept

The proposed approach is to provide the methodological basis for a system able to perform fully autonomous navigation. It is proposed to investigate the combined use of vision, landmarks, and triangulation, to automatically estimate robot position and heading during robot navigation. The following sub-sections will present proposition and argumentation for the solution model.

As discussed in section 1.7, the main challenges of the proposed combination are represented: localization error, landmark recognition and automatic model acquisition and localization. The investigated system concept represent an attempt to address and possibly alleviate these challenges.

3.1.1 Choice for World Sensing: Vision

The main sensory modality chosen for the robot localization is Vision which is believed to posses a great unexploited potential.

Vision is the dominant sensor modality for humans and animals who successfully exploit it to "drive" their bodies during any navigation task. Vision in humans works quickly and effectively. Without the assistance of vision, the humans navigation capability is dramatically reduced to very simple tasks and much slower. For example, even people who have been blind since birth, and who have developed enormously other sensor modalities, are not able to perform navigation in the same efficient way as the people who can see. In addition to its potential, vision has been chosen since it can provide a rich and discriminant information, i.e. the environment visual appearance, which can be exploited as natural landmark. The visual appearance may in fact contain objects or part of them naturally occurring in an environment.

Unfortunately, despite the great trust on its potentiality, vision-based technology is not mature yet to allow a reliable recognition. Vision systems are sensitive to changes in illumination conditions and only few robust algorithms exist which typically are computational expensive and/or require a consistent amount of human assistance to operate.

The task of the vision system in the proposed method is to identify landmarks in the camera image-plane. In particular, as explained in section 1.7, the vision system must identify three landmarks. Every time a triplet of landmarks has been identified in the image-plane, the robot position and heading can be estimated by a triangulation method if landmark world position is known.

The operation of identifying three landmarks in the image plane may be computational expensive, depending on the system and landmark characteristics. This leads to the fact that the robot would need to either, stop during navigation, (and re-orient itself at each run of pose computation), drive "blind" for some time, or adopt a continuous pose updating scheme based on internal sensors while a new absolute robot pose is being processed. The latter would allow the system to keep track of the current robot pose until when a new pose estimate, based on triangulation, will become available. The continuous updating scheme is the preferred solution, in particular, the use of *odometry* is proposed to support vision.

As discussed in section 1.5.1. The odometric system is available on most of robotic platforms and it allows for a reliable pose updating on the short term of navigation. This makes acceptable the choice of relying on odometry in case it is a short distance the one traveled by the robot between two triangulation-based pose estimates. This way of updating robot position is analogous to what proposed for outdoor GPS[1]-based navigation systems, where the GPS estimate is provided at certain intervals, forcing the system to either drive "blind" or rely on some other type of sensor, between GPS readings.

In summary, it is proposed the use of vision assisted by odometry.

3.1.2 Choice for Pose Estimation: Triangulation

This subsection represents the proposed solution to the issue of *localization error* mentioned in section 1.7, as well as the answer to the question: how should the Vision output be used?

The proposed system is intended to be used in indoor semi structured environments. Consequently, the robot is only moving around on a planar surface, which makes sufficient that the robot localizes itself by estimating its 2D position and heading.

There are different ways to use landmarks for robot self-positioning. A triangulation based localization can in general be achieved by measuring the angle the camera needs to be rotated to fixate the relevant landmarks. In particular, by measuring three angles and knowing where the corresponding landmarks are, the position and heading of the robot may be derived. Figure 3.1 demonstrates the monocular triangulation. As shown in the figure at least three

[1] The term GPS stands for *Global Positioning System*, see section 1.5.2

angles are required, since there are three unknowns to be determined, i.e. X,Y, and θ.

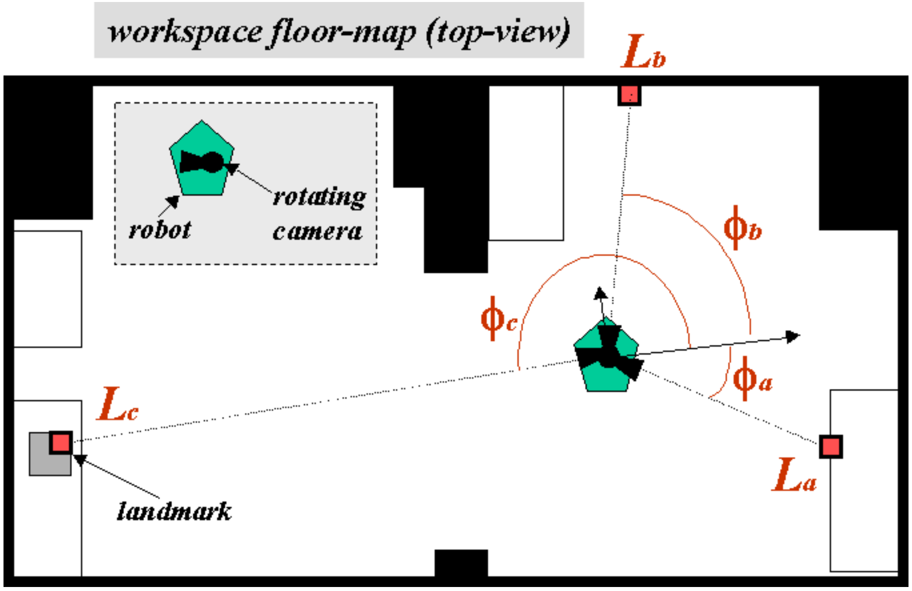


Figure 3.1: The figure demonstrates the monocular triangulation. By knowing three angles and knowing where the corresponding landmarks are, the pose of the robot may be derived.

Instead of measuring the angle to different landmarks the distance to these could be used. This method is known as *Trilateration*, whereas the use of angles is denoted as *Triangulation*. However, because of the problem with acquiring accurate range estimates from vision, particularly in cluttered scenes, angular measurements are preferred. The angular measurement is known to be very robust and comparatively easier to achieve. Since a single measurement is obtained from each landmark, at least three landmarks are needed.

But why is it that three landmarks are needed?

For example a triangulation scheme as that used for nautical and flight navigation may be done from two known landmarks (beacons). In this case, however, one additional known parameter, that is the global orientation of the vehicle, is estimated using a compass. The vehicle can then use the heading or distance to two landmarks to perform self-positioning.

An alternative to triangulation would be to only rely on one landmark and use the angular separation between the predicted and the observed line of sight to the landmark, as an input to the pose-computation procedure. Then, from this angle, a new robot position can be computed by using a Kalman filter framework as presented by Crowley, [36]. Figure 3.2 shows a schematic representation of the mentioned pose computation schemes. A drawback to this alternative method is that robot position can only be computed relative to a previous estimate, which makes the system strongly relying on, for example, odometric measurements.

The conclusion is that having a controllable camera on a robot which navigates a typical indoor environment it is quite straight forward to consider the triangulation technique based on at least three independent measurements, i.e the angle to each landmark. In this way it is imagined that the localization process will have the characteristic of being fault-tolerant.

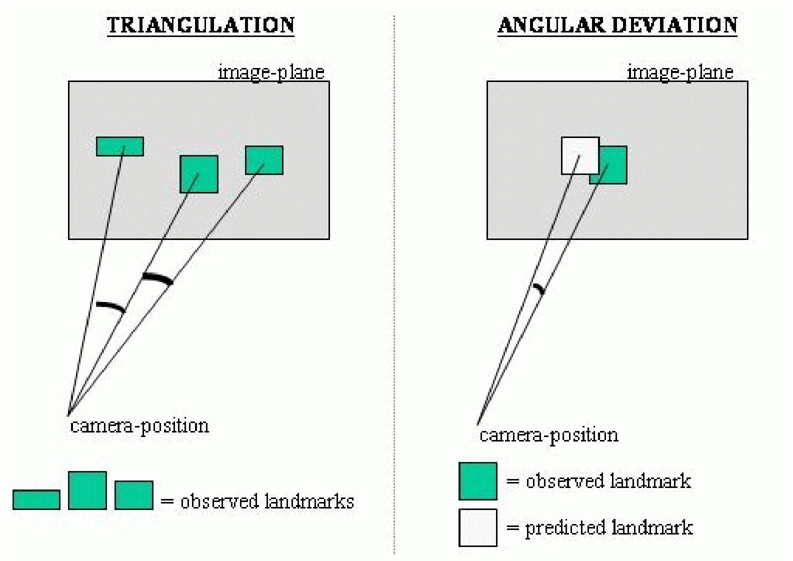


Figure 3.2: The figure shows a schematic representation of both triangulation and angular deviation.

A new *absolute* position estimate is in fact provided at each run.

A number of independent angular measurements higher than three can also be used (Greiner and Isukapalli, [62]). This is likely to improve accuracy and reliability of the estimated positional information. However, the use of only three landmarks, (i.e. the minimum number), other than allowing for a faster detection (see section 4.1), would make it easier to have all the needed landmarks in the camera field of view, (i.e. all the landmarks can be detected in the same image), without a very high number of landmarks in the environment. The use of a sparse set of landmarks represent an advantage because it will simplify the landmark acquisition phase, and because in some type of indoor environment, (e.g. white walls corridors), there may be a limited number of distinctive landmarks, (an important characteristic for the method proposed in this thesis), which can be extracted from images.

An alternative would be to observe the needed landmarks on different images. The camera could pan in the direction of the landmark and capture an image every time a landmark has to be observed. However, this procedure is not advisable since positional errors may arise due to the robot movement elapsed during the different image acquisition.

In this work it is consequently chosen to study the use of exactly three landmarks (a landmark triplet) for triangulation and to consider the case when they are all visible from the same camera field of view. The focus is on how accurately the robot pose can be estimated using this approach and on how much there is a need for further landmarks to be taken into account. In summary, the use of only three landmarks will test limits and potentiality of the approach.

3.1.3 Choice for World Representation: Natural Landmarks

This subsection represents the proposed solution to the issue of *landmark recognition* mentioned in section 1.7, as well as the answer to the question: what shall the Vision modality "see"?

The Vision system is responsible for "seeing" the landmarks needed to estimate robot position and heading. In particular, the vision system estimates current landmark locations in the image plane. To achieve this result, the system has to be able to recognize landmarks previously acquired in the incoming views.

But what type of landmark best suits the system requirements?

The landmark types analyzed in the following paragraphs represent characteristics of the environment which can be detected in images. There are different paradigms adopted in the literature concerning recognition of environment characteristics in the image plane. It is then proposed to categorize the possible types of landmarks into three level of abstraction. The higher the abstraction the more this is remote from the image domain.

At the lowest level there is the representation of environment characteristics which is the closest to the sensor input and so closely tied to imaging process. This level may be called *appearance based*. A higher level is called *feature based*, whereas the highest level may be called *model based*.

Appearance-based

An *appearance based* landmark is directly described by the raw sensor input, i.e. by the light intensity values contained in the image plane. A landmark could then be an image of the workspace room, or a sub-image representing a certain portion of it. When a landmark is extracted from an entire image, the selection might be based on certain patterns or values.

Appearance based landmarks contain information close to the sensor readings, so very discriminant. The characteristic of being discriminant means to have a rich, high-contrasted and unambiguous appearance, so that a landmark can be unique inside a workspace. On the other hand, this information is not invariant to most of the parameters governing the imaging process: lighting, viewpoint, intrinsic camera parameters, etc.

The appearance based representation is a very attractive option, since no error-prone data transformation occurs and a discriminant information is available. However, the recognition of appearance landmarks is based on template matching techniques, which typically is computational expensive. In addition, appearance is a non-invariant representation, so that the use of this type of landmarks would imply the need for many different appearance-based landmarks to be acquired, maintained and matched during navigation. In other terms, a great deal of memory and computational resources.

Feature-based

In order to provide a more general characterization of the environment, the sensory input could be processed to extract *features*. For example, a geometric feature such as a vertical edge or a corner. A feature could also be a color dominating an image, a symmetry propriety, etc. Features represent a level of abstraction higher than appearances since the "meaning" of a feature can be extrapolated from the specific context. This allows for viewpoint invariance as well as a reduction of the amount of data to be associated to a landmark.

A *features based* matching process is much faster than an appearance based, since the amount of information to be compared is much lower. The lower amount, however, also indicates a less discriminant information which calls for additional information to be associated to features, such as, the distance between features, the position of features in the environment, etc.

Model-based

The highest level of abstraction in the proposed categorization is called *model based.* At this level features and/or appearances are combined by an image process which is more remote from the image domain than the other two paradigms. A model can represent a combination of geometric features, such as the case of a CAD model, as well as a combination of appearances, such as the case when different views of a certain object are combined into one representative image.

A model can also combine features and appearances among them as well associate them to some amount of 3D information, such as measurements, positions, etc. For example, a door can be characterized by a collection of characteristics such as a wide flat surface containing a generalized slim cylinder at certain height above the floor, (i.e. a handle), plus the door picture and its position in the environment. A model can even be a concept. For example, a door can be characterized by the concept that it can be opened or that the robot has been able to run through it.

It becomes then quite obvious the fact that building up a model can be a complex operation. Furthermore, the world is quite often too complex to be totally modeled which leads to approximated models. An approximated model also means a potential loss of information while errors may arise as well. A model representation has, however, the great advantage to be the most invariant since it is the most removed from image processing. As for being discriminant, it depends on the model type, complexity and level of approximation.

A *model-based* recognition of landmarks may involve different types of operations depending on the model and applied method. For example, a recognition task for a door can be, detecting that it is being opened by analyzing changes in surface reflections, as well as, in case of a geometric model, matching the edges contouring the door.

Proposed Solution

Figure 3.3 summarizes main characteristics of the abstraction levels described above. As we can see the characteristics for a landmark of being discriminant, invariant, and computational expensive, depend on the chosen representation paradigm. The chosen paradigm influences the level of abstraction for the matching process, which can therefore take place at appearance, feature or model level. The higher the abstraction the more complex is the transformation to apply to the input image to allow a match. Figure 3.4 show all the possibilities for a match to take place and the required transformations.

Traditionally, if model representation is too removed from imaging process, robust matching becomes difficult (appearance based paradigms has been extremely successful over the last

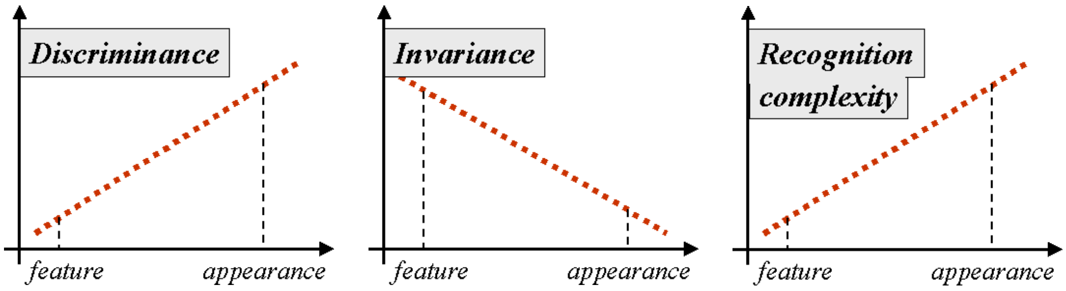


Figure 3.3: The figures summarize main characteristics of the described abstraction levels.

decade). On the other hand, if model representation is entirely appearance based we have viewpoint invariance problem.

This becomes quite immediate that to just rely on appearances, or on features, (always intended as simple general image characteristics), will not be a satisfying solution. There is consequently a need to propose something at model level, making sure that a proposed model possesses the following characteristics:

1. simple, i.e. a coarse model (not dense);
2. closely related to imaging process;
3. containing additional information in order to generalize the raw camera input, i.e., to make it more invariant.

The solution which is proposed in this thesis intend to combine the advantages of different paradigms. It is proposed to rely on the appearance of *medium-sized landmarks* (i.e. substantially smaller than an image), and on landmark positional information (landmark 3D position and orientation). The proposed landmarks are *natural visual landmarks*, that is, a structure in the environment which can be recognized in images. Consequently, the visual landmark is merely a sub-image of an observed view.

A popular choice is naturally occurring, visual landmarks which represent planar textures such as door signs, posters, light switches, etc., but also just a part of objects. The exploitation of such texture information, even for a small sized texture, provides a natural and discriminant information, (see figure 3.5).

Besides being discriminant, a visual landmark is required to allow for its recognition in presence of changes in viewpoint observation. The landmark image-resolution consequently becomes an important matter. A larger texture-patch is usually more discriminant than a smaller one, since it represents a wider workspace portion so that many objects could be represented inside of it.

On the other hand, a texture-patch representing many objects can easily change its appearance when the viewpoint changes because of occlusions, and changing viewing angle. This reduces the possibility for a landmark to be recognized in presence of changes in viewpoint observation. Furthermore, small landmarks allow for faster image-processing.

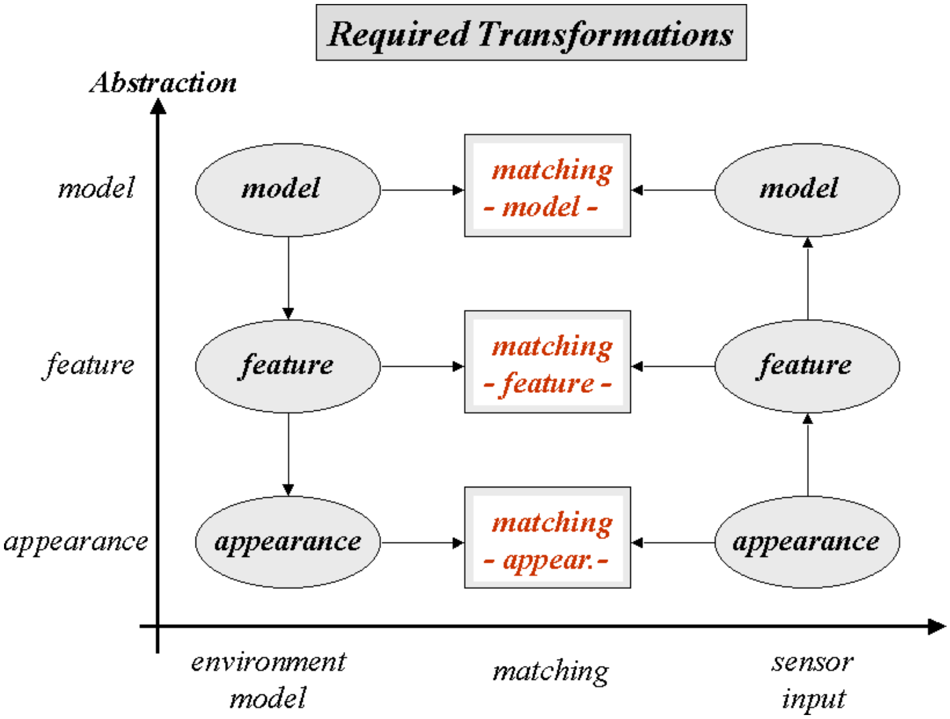


Figure 3.4: The figure shows all the possibilities for a match to take place and the required transformations.

The problem of finding a suitable size for landmarks is an unsolved problem as it can be seen from the many proposed landmark dimensions, (Madsen and Andersen [97], Zingaretti and Carbonaro [152], Trahanias et al. [141], Balkenius [9], Davison and Murray [41]), and it appears that an optimal size is very depending on the specific approach used for the localization. The author of this thesis has investigated this problem as well, (Livatino and Madsen, [93], [92], [94]). A satisfying compromise for typical indoor environments which fits the proposed self-localization method has been found through simulations and experiments with a real system. An image of 32x32 pixels represents a good compromise when a 8mm camera focal-length is used and image-resolution is 512x512 pixels. Figure 3.5 shows examples of proposed landmark texture-patches. The proposed size for visual landmarks represents a "medium-sized" texture patch if compared to other types proposed in the literature, (see chapter 2).

The appearance of a medium-sized landmark will allow for a reliable recognition based on template matching techniques, whereas the position of the landmark will facilitate back projection of landmark appearance in the image plane from arbitrary viewpoints.

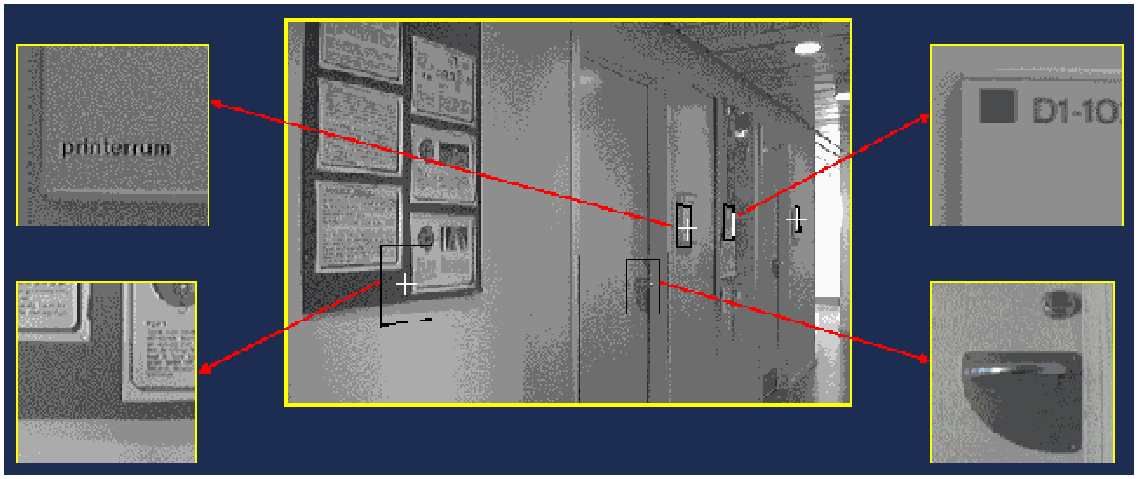


Figure 3.5: Examples of proposed visual landmarks.

3.1.4 Additional Choices

Monocular Vision System

The proposed vision system is monocular. In particular, there is one camera on a pan unit sitting on top of the robotic system at the height of 1 meter and able to freely pan the surrounding environment over 360 degrees. This solution has been considered sufficient for the proposed learning and localization in typical indoor environments, and so preferred to the one using a stereo head. In particular, the possibility provided by stereo camera head to compute range measurements to landmarks represents an option not required for the proposed localization, since it is preferred the more reliable solution of estimating the angle to a landmark, (see section 3.1.2).

As for the learning phase, the accuracy required in landmark pose estimation can only be satisfied by a large baseline stereo configuration. For typical distances to landmarks in indoor environments, a baseline of around one meter represents a reasonable tradeoff. Such a baseline would however require an impractical large camera head for the robot, so a solution has been preferred where the large baseline situation arises when a landmark is identified in two monocular observation taken from different positions, (a technique sometime referred as *Active Triangulation*). The monocular solution also allows for a wider view-field of observation, not limited by the presence of a close coupled camera. This advantage would allow in a future the use of omni-directional cameras (with a high resolution). In conclusion, a monocular vision solution represents a suitable setup to experiment and test performance. Figure 3.6 represents an example of monocular camera head (the one used for the experimentations presented in next chapters).

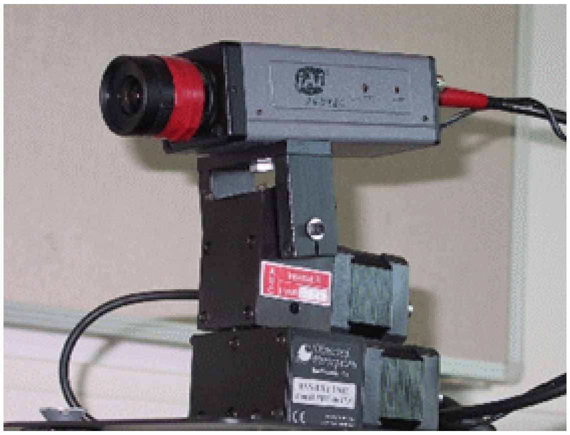


Figure 3.6: The figure represents an example of monocular camera head (the one used for the experimentations presented in next chapters).

Optimal Selection of Landmark Triplets

Naturally, when using real data the detection of landmark image locations in the image-plane will be error prone. Madsen and Andersen, [97], show that under good conditions template matching (normalized cross correlation) can determine image locations of landmarks with a standard deviation ranging from approx. 1.2 pixels to 3.6 pixels and an average of 2 pixels. The effect of this error on the computed position, relative to a floor map, can be visually illustrated using the simulated example of figure 3.7.

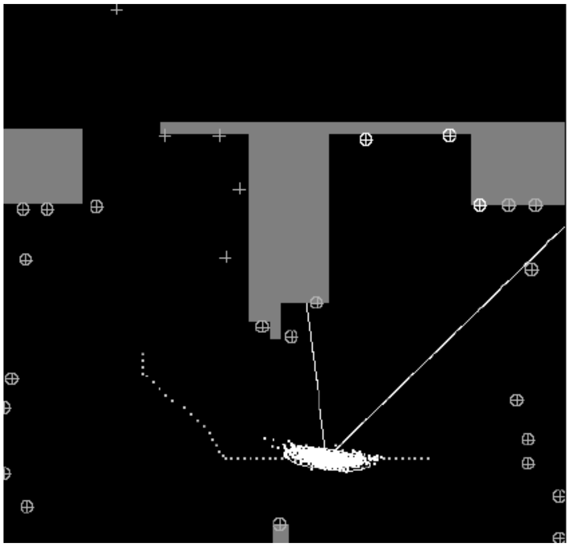


Figure 3.7: 2D floor map of robot workspace. Obstacles are shown in grey, free-space in black. The robot is following an assigned path, represented by small white points. All landmarks are shown as crosses. Visible landmarks are within a circle and those used for computing position have a white circle. The cloud of white points are robot positions computed with noise added to the sensory measurements. The two white lines emanating from the robot position indicate the camera field-of-view. The ellipse is the positional uncertainty computed by the 1000 simulated runs, (Livatino and Madsen, [93]).

Figure 3.8 top-row shows main parameters involved in optimal triplet selection. The ellipse in the figure represents the expected positional uncertainty in robot pose associated to the selected triplet. The optimal triplet is the one which minimizes such an uncertainty, (see subsection 4.2.3). Figures 3.8 bottom-row illustrate a synthetic example representing how the noise sensitivity may vary drastically with the choice of landmarks. This indicates the need for a strategy for selecting *good* landmark triplets, which in turn requires a technique for evaluating the possible landmark triplets before using them for triangulation.

The proposed solution is a simple strategy for choosing the best landmark triplet, if more triplets are available, by predicting positional uncertainty. In particular, the system can choose the landmark triplet which minimizes the noise sensitivity. The selection of landmark triplets is a fast and automatic process. The advantages of the proposed method is an accurate pose estimation, since the optimal triplet is always selected. It has to be noticed that a previous estimate of robot pose it is always assumed.

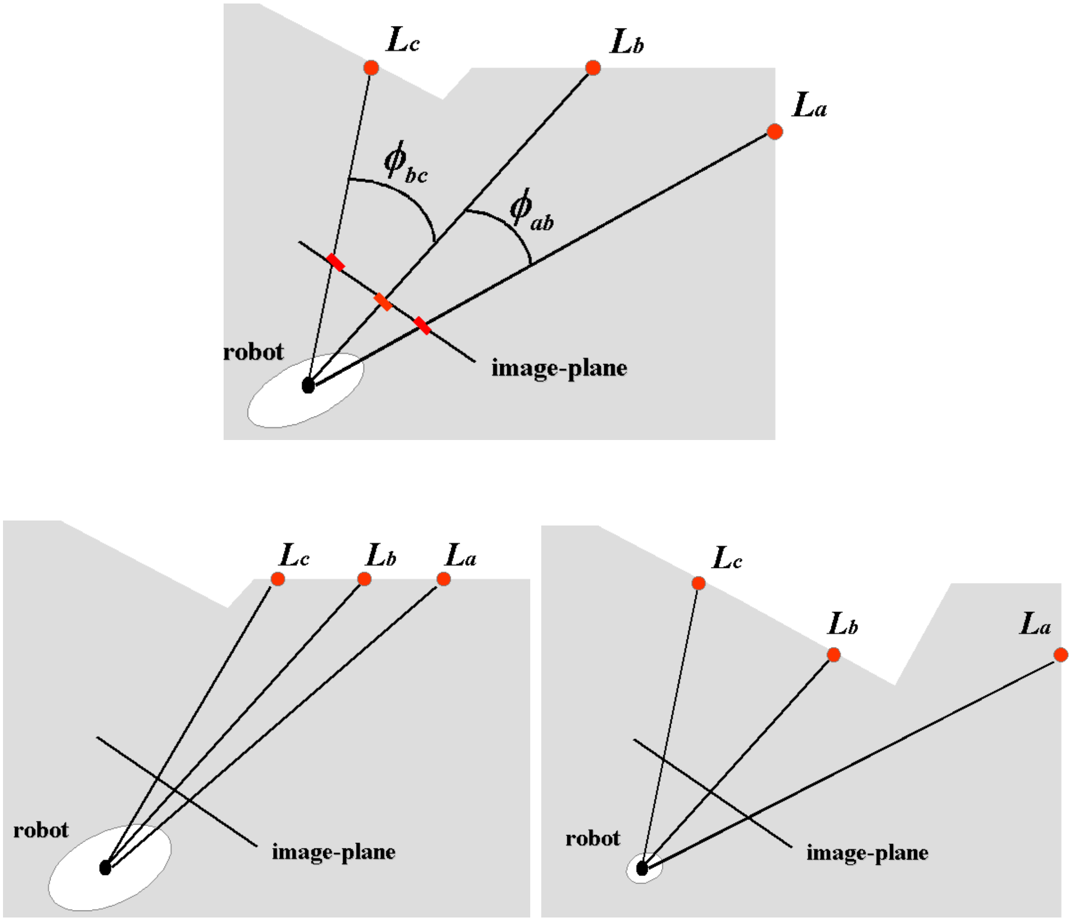


Figure 3.8: The figure top-row represents main parameters involved in optimal triplet selection. These are denoted by: red dots (landmark position), red short segments (landmark location in image-plane), the white ellipse (the expected positional uncertainty in robot pose). The optimal triplet is the one which minimizes the ellipse area. The figures bottom-row illustrate synthetic examples representing variation in noise sensitivity when choosing different landmark triplets, but keeping the same robot position. In the bottom-right figure the chosen triplet provides a much better position estimate than that in the left figure.

3.1.5 Choice for Autonomous Behavior

This subsection represents the proposed solution to the issue of *automatic model acquisition and localization* mentioned in section 1.7. The choice for pose estimation represent the answer to the question: how should the vision modality "behave"?

A fully autonomous behavior is represented by:

1. the possibility of self-localization, i.e. automatic estimation of robot pose during navigation;
2. automatic learning, i.e automatic acquisition of the necessary information to perform the localization task.

As introduced in subsection 1.5.4 and chapter 2, there are different navigation schemes for autonomous navigation. Mainly, "2-phases" (learning and localization) or "SLAM" (simultaneous localization and mapping). For now, we consider the "2-phases" solution. Figure 3.9 illustrates the proposed "2-phases" navigation strategy. Consideration for a "SLAM" conversion to be addressed in future work.

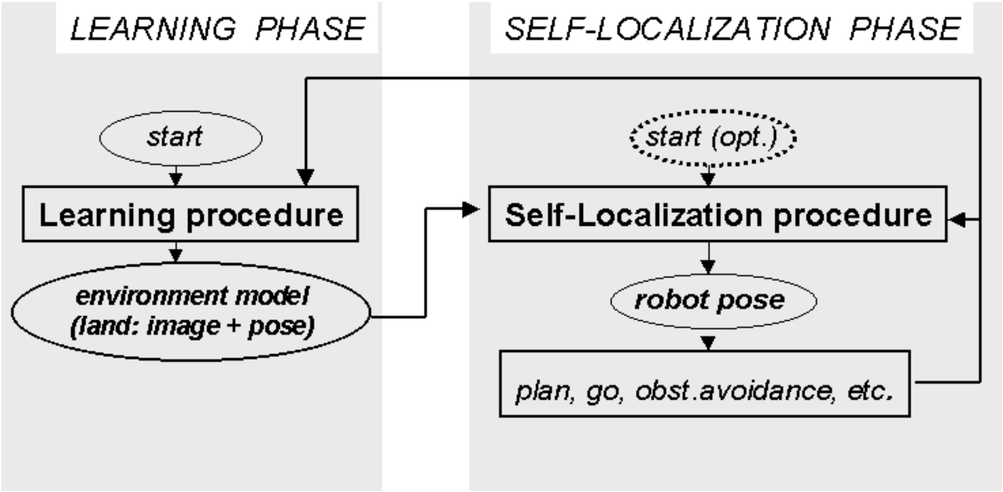


Figure 3.9: The figure illustrates the proposed "2-phases" navigation strategy: Learning-phase, i.e. automatic acquisition of visual landmarks; and Self-Localization phase, i.e. estimation of robot pose during navigation.

Self Localization

The sensory input to the localization method is an estimate of the current robot pose and an image of the workspace expected to contain the optimal triplet of landmarks. The prior knowledge is represented by camera parameters and environment model, which is represented by landmark representative images and landmark poses in the environment. The output is the robot position and heading.

Automatic localization is only possible if the error in robot position during navigation is small enough to allow recognition of a triplet of landmarks in the image plane. The pose estimate is more accurate if its computation is made more frequently. However, a faster landmark-recognition process also means a less precise procedure (leading to a higher error). The problem of positioning and updating speed thus become coupled.

It is difficult to find a general tradeoff between estimated accuracy and updating speed because of the non-systematic error which affects the sensory system. In particular, the error in the odometric system is unpredictable and depends on the followed trajectory.

A compromise solution is proposed where the landmark recognition is based on a computational expensive but accurate recognition of landmarks in the image plane (based on normalized cross correlation). This solution has been preferred to other faster alternatives because of the reliability shown in experiments for typical indoor trajectories.

The accuracy provided by the proposed localization algorithm (based on triangulation), allows the system to set the odometry to the new estimate every time a landmark triplet is recognized. This prevents errors from accumulating, hence, the system can run autonomously for a long time.

The method proposed for self-localization is thoroughly described in chapter 4. Figure 3.10 shows a schematic representation of the proposed localization process. There are three main steps of computation: Landmark Selection, Landmark Detection, Pose Computation.

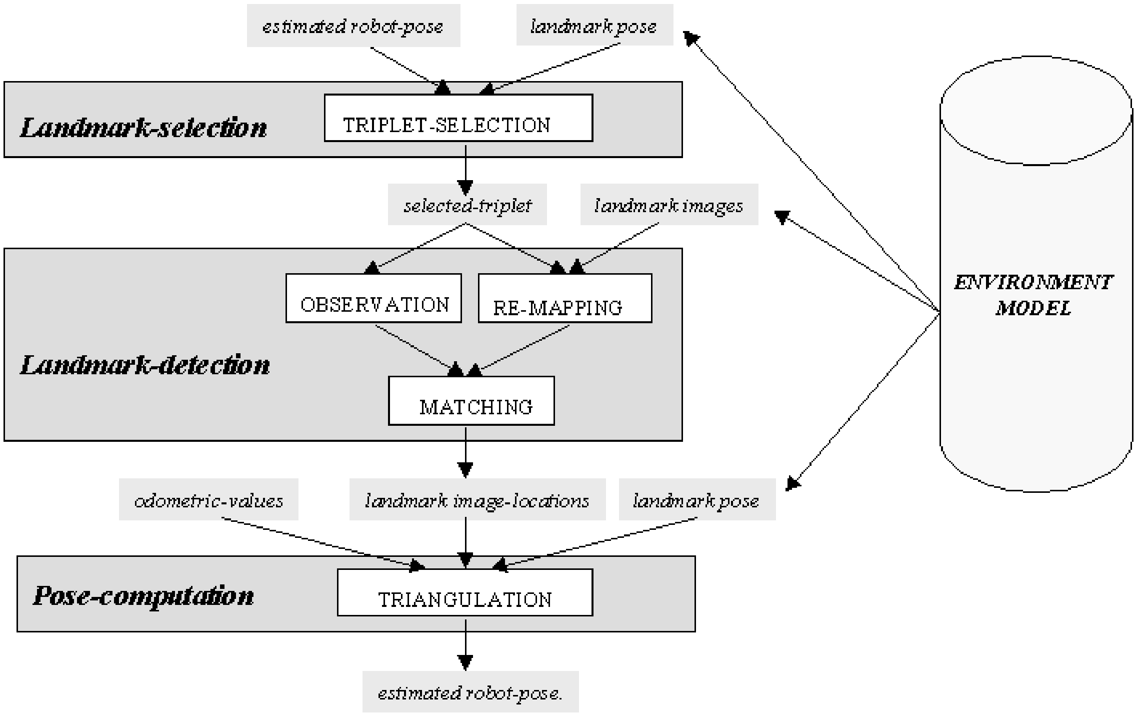


Figure 3.10: A schematic representation of the proposed self-localization procedure. The figure represents main computational steps and required parameters.

Automatic Learning

The sensory input to the learning method is a collection of images of the environment surrounding the robot, and an estimate of the current robot position. The latter can be unknown for the very first position, (and could be considered the world frame origin). The prior knowledge is the camera parameters. The output is representative images of the landmarks and their pose in the environment.

Automatic localization is only possible if the error in landmark appearance and pose is small enough to allow recognition of a learned triplet of landmarks in the image plane. This method for automatic learning must be designed in order to allow for "accurate enough" visual and positional information. In summary, the method for automatic learning must be able to: (1) automatically acquire landmark representative images; (2) automatically estimate landmark pose in the environment; (3) allow for "accurate" visual and positional information.

In order to successfully operate, the proposed automatic learning needs to know robot position and heading above a certain accuracy (see chapter 4). If this requirement is satisfied, the proposed learning can be executed both before and during navigation.

It is proposed to extract landmark representative views from high-resolution stereo images of the environment, and estimate landmark poses based on a large baseline stereo matching. The method proposed for automatic learning is thoroughly described in chapter 5. The proposed method is imagined capable of providing discriminant and invariant landmarks, as well as accurate landmark pose estimates, expected to allow for their automatic recognition during robot navigation. The method proposed for automatic recognition is thoroughly described in chapter 6. Figure 3.11 shows a schematic representation of the proposed learning process. There are two main steps of computation: Landmark View Acquisition and Landmark Pose Computation.

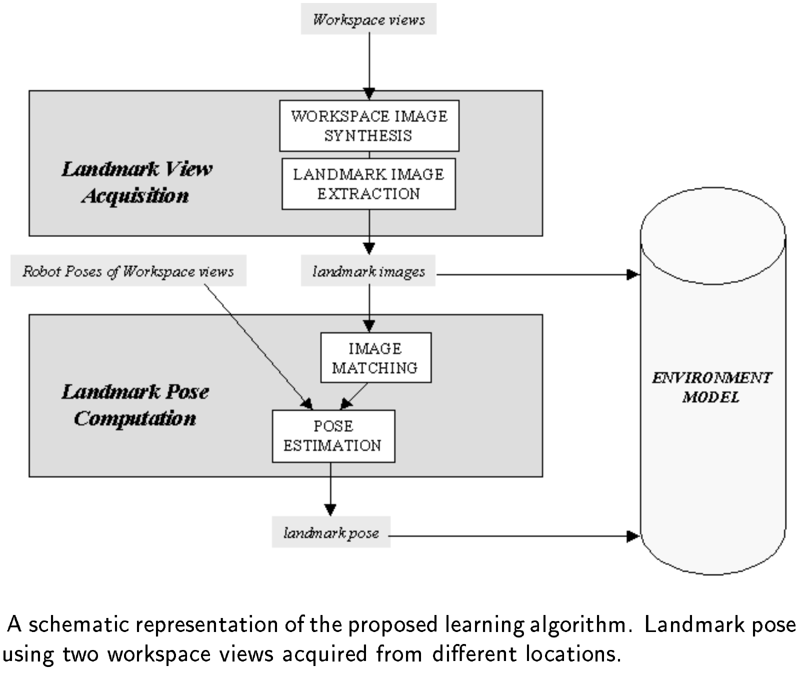

Workspace views
WORKSPACE IMAGE SYNTHESIS
LANDMARK IMAGE EXTRACTION
Landmark View Acquisition
Robot Poses of Workspace views
landmark images
IMAGE MATCHING
POSE ESTIMATION
Landmark Pose Computation
landmark pose
ENVIRONMENT MODEL

Figure 3.11: A schematic representation of the proposed learning algorithm. Landmark pose computation is performed using two workspace views acquired from different locations.

3.2 System Demo

The aim of this section is to show the reader what the system ended up being by providing an overview of the main system functionalities at a glance. In particular, the proposed system concept is represented through a *system demo*, that is, a ”visual description” of consecutive processing steps that the system needs to go through in order to perform an autonomous and accurate robot localization.

The main system functionalities, illustrated in the following paragraphs, can be listed as: (1) panoramic view synthesis, (2) landmark candidate extraction, (3) landmark pose computation, (4) landmark model memorization, for the learning phase; and: (5) optimal landmark-triplet selection, (6) landmark visual prediction, (7) landmark detection, (8) robot pose computation, for the localization phase.

(1) Panoramic View Synthesis

The first action to be taken by the system during the automatic learning phase is to acquire and synthesize a few panoramic views of the environment surrounding the robot. This action can be performed from any arbitrary workspace position. Figure 3.12 represents a typical synthesized panoramic-view from our experiments. The view is based on 72 high resolution images acquired by panning the monocular camera system over 360 degrees, and then merging the acquired views using cylindrical projections, (see subsection 5.1).

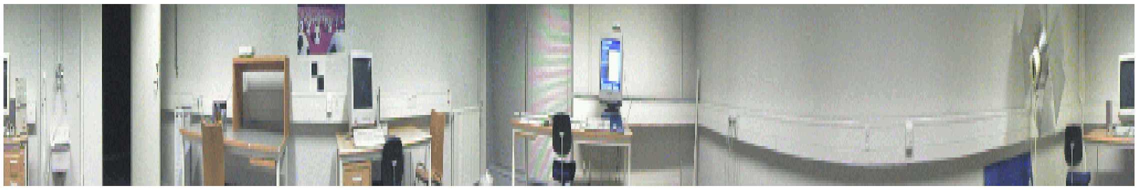


Figure 3.12: The figure shows an example of a synthesized panoramic view.

(2) Landmark Candidate Extraction

Once a panoramic view has been synthesized a technique is proposed to process the view in order to extract landmark candidates. These are image regions which represents distinctive workspace portions. Figure 3.13 shows a typical set of extracted texture patches representing landmark candidates. The landmarks are extracted by means of an attention selection mechanism, (see subsection 5.1.3).

(3) Landmark Pose Computation

Once a set of landmark candidates has been extracted, their *pose* is estimated. The landmark pose is defined as 3D position of the object represented in the landmark view and the orientation of landmark surface (landmarks represent planar surfaces).

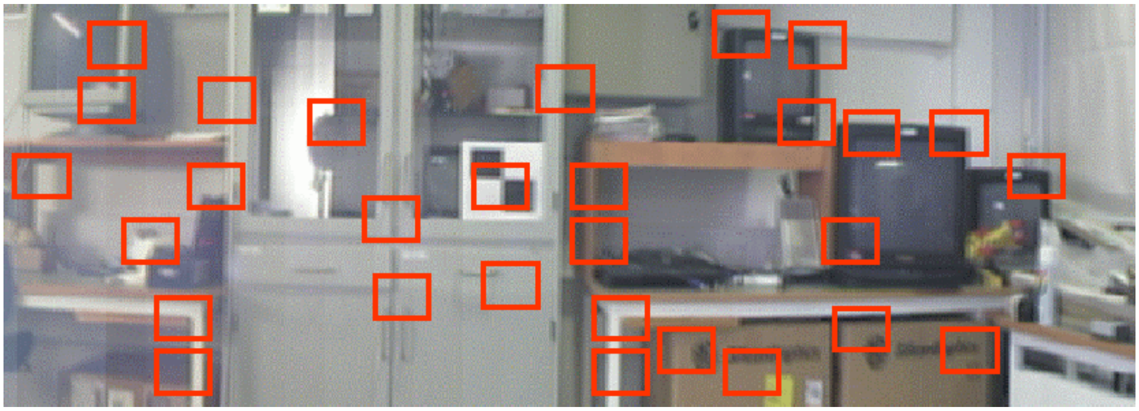


Figure 3.13: The figure shows an example of selected landmark candidates.

The computation of landmark pose is achieved by first estimating correspondence between the same landmark candidates in different panoramic views, and by then estimating position of landmark centers and extremes by a *stereo triangulation* technique, (see section 5.2). Figure 3.14 illustrates the stereo reconstruction process, (cylindrical panoramic views and corresponding texture patches).

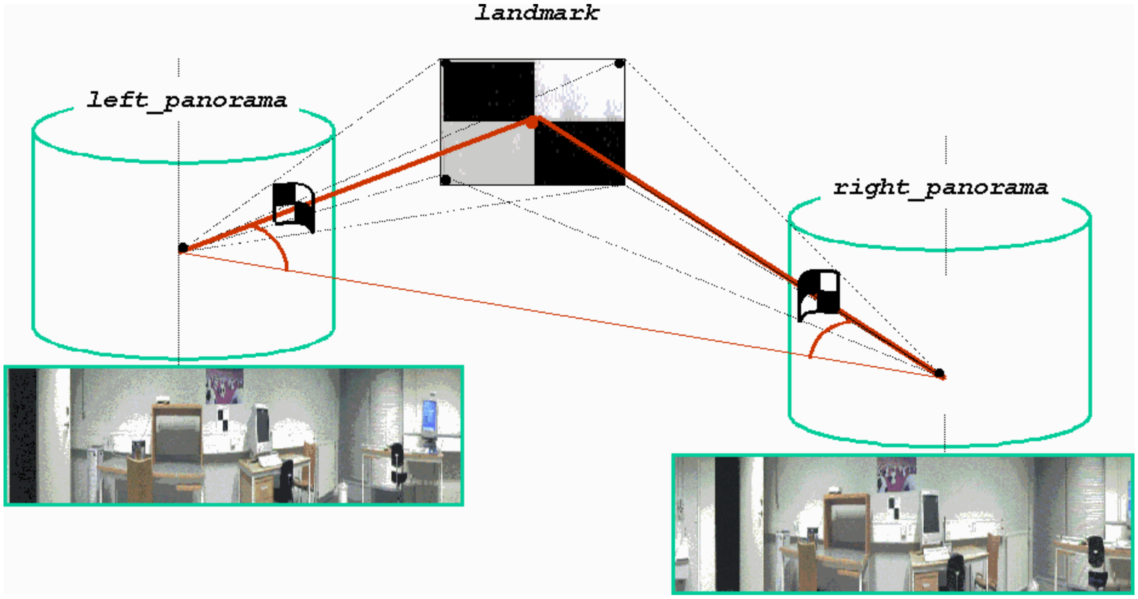


Figure 3.14: The figure shows the main factors involved in *stereo triangulation*.

(4) Landmark Model Memorization

Landmark reference views are subsequently stored in a database together with the estimated positional information, to be used by the system during the self-localization phase. Figure 3.15 shows an example of a stored landmark set (environment model).

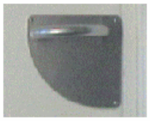
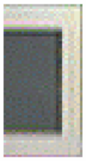
Ref. view			
Pose			
X	*78.0*	*5322.0*	*481.0*
Y	*2330.0*	*5009.0*	*1726.0*
Z	*1603.0*	*1054.0*	*1071.0*
θ	*90.0*	*0.0*	*85.0*

Figure 3.15: The figure represents an example of a stored landmark set to be used during robot navigation for self-localization.

(5) Optimal Landmark-Triplet Selection

The first action to be taken by the system during navigation in order to localize itself, is the selection of the *optimal landmark triplet*, (relative to current robot position). That is, the triplet which minimizes positional uncertainty in robot pose, given the set of all visible landmarks available at the current robot position.

Figure 3.16 shows variation in noise sensitivity when choosing different landmark triplets, relative to the same robot position. The cloud of white points are robot positions computed with noise added to the sensory measurements. The optimal landmark triplet is that one which minimizes the noise sensitivity. Figure 3.17 shows the selected optimal landmark triplets along a certain path.

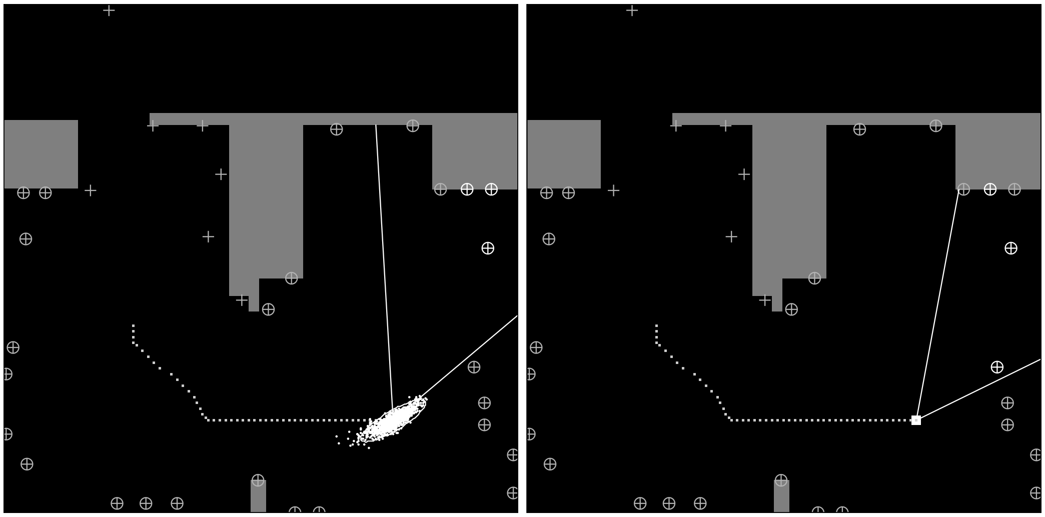


Figure 3.16: Variation in noise sensitivity when choosing different landmark triplets, but keeping the same robot position. In the right hand figure the chosen triplet provides a much better position estimate than that in the left figure. The figures represent 2D floor maps of robot workspace. Obstacles are shown in grey, free-space in black. The robot is following an assigned path, represented by small white points. All landmarks are shown as crosses. Visible landmarks are within a circle and those used for computing position have a white circle. The cloud of white points are robot positions computed with noise added to the sensory measurements. The two white lines emanating from the robot position indicate the camera field-of-view. The ellipse is the positional uncertainty computed by the 1000 simulated runs, (Livatino and Madsen, [93]).

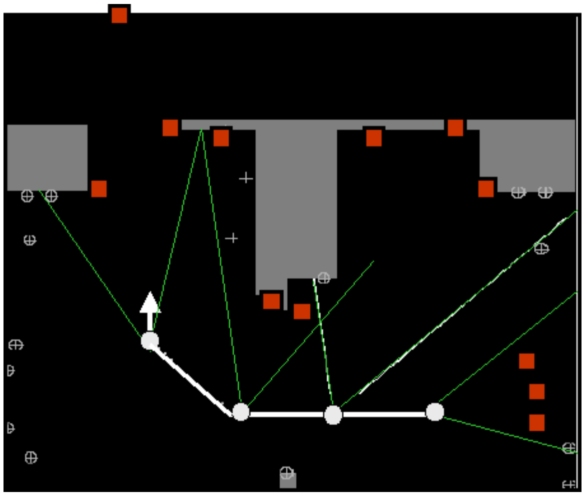


Figure 3.17: The figure shows a typical path on the laboratory floor-map. The given path represented a trajectory that robot had to run along in order to get the laboratory exit-door. The green lines emanating from robot positions (white circles) represent the camera field of view. The little red squares represent selected landmarks.

(6) Landmark Visual Prediction

Once three landmarks have been selected, their reference views, (previously learned and stored in a database), can be re-mapped to how they should appear from current robot pose. This operation is performed by a proposed method for *realistic virtual-view synthesis*, (see section 6.3). These views can then be used as prediction in a template matching process. The generated landmark re-mapped views are consequently named *visual prediction*.

Figure 3.18 represent the process of creating a realistic virtual-view. The figure visually describes the transfer of texture values from cylindrical reference-views to the visual prediction for an example landmark. The image top-right represents the "current observation". The visual prediction represents the view re-mapped from reference images according to current camera viewpoint.

Figure 3.19 show examples of landmark visual predictions. The figure right-hand and bottom left represent landmark generated *visual predictions* from different viewpoints. For comparison, observed landmark views are showed in figure top left.

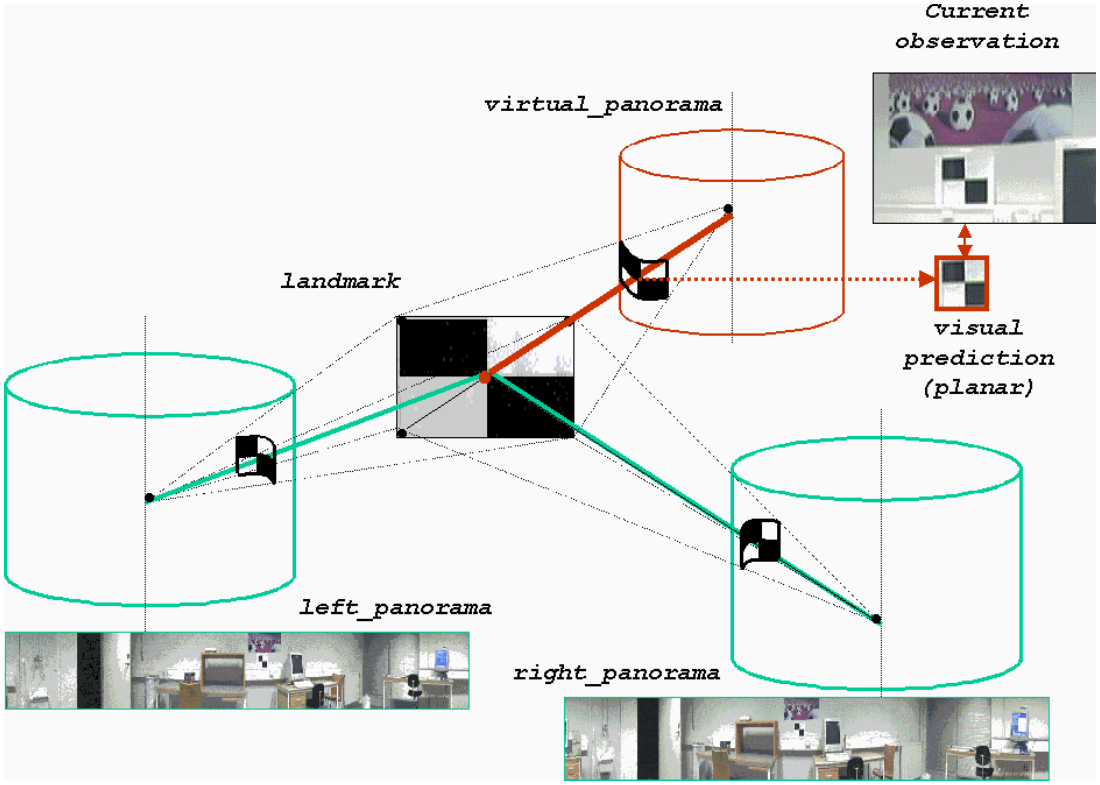


Figure 3.18: The figure represent the process of creating a realistic virtual-view (see text).

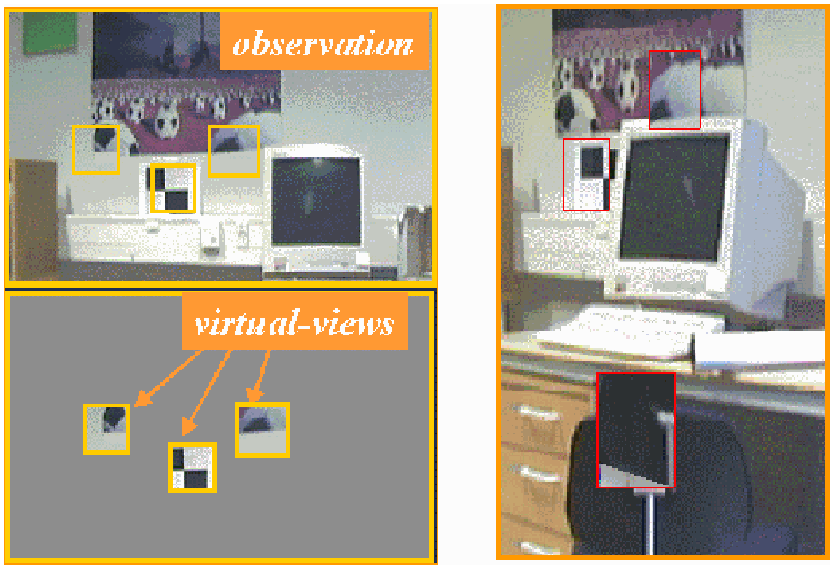


Figure 3.19: The figure shows examples of landmark visual predictions (see text).

(7) Landmark Detection

The landmark detection process has the goal to locate selected landmark triplets in the currently observed view. The proposed method is based on a template matching technique which compares current views and re-mapped views (visual prediction), within an estimated "search window".

Figure 3.20 shows a schematic representation of the re-mapping and matching process.

(8) Robot Pose Computation

The detected location of the landmark in the image-plane represents the input to the triangulation method which is then used for computing the camera position and heading, which in turn can be converted to robot position and heading, (see subsection 4.2.1).

Figure 3.21 represents the triangulation process.

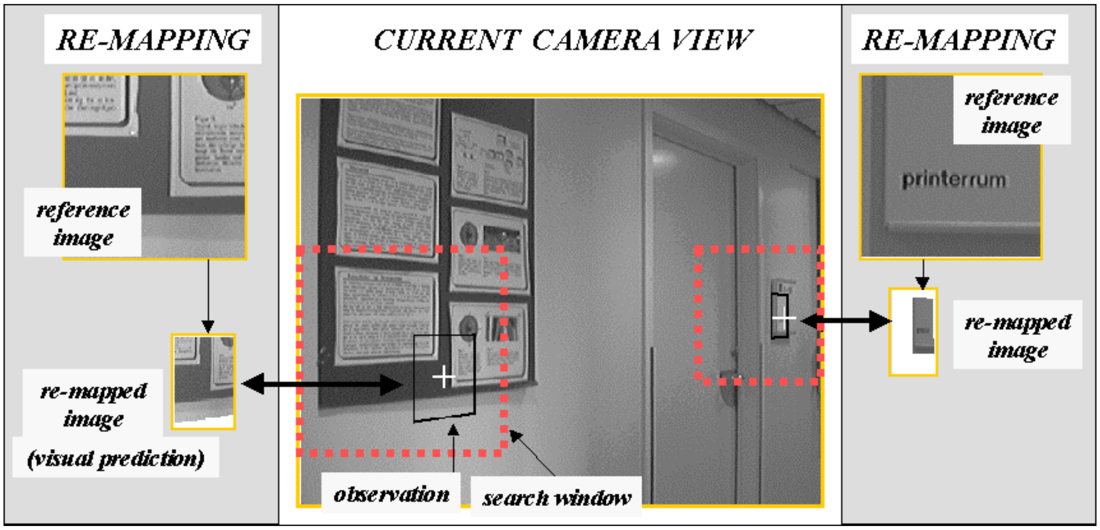


Figure 3.20: The figure represents a schematic representation of the landmark re-mapping and matching process.

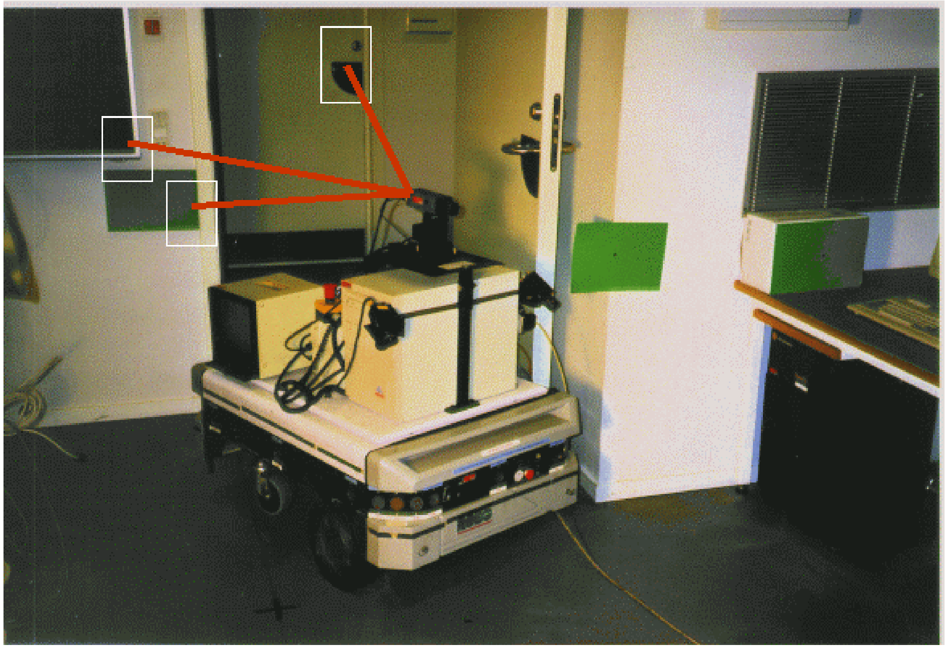


Figure 3.21: The figure represents the triangulation process. The three red lines emanating from the on-board camera symbolize the triangulation process. In particular, these lines connect each landmark center to the optical center of the camera.

3.3 Approach Challenges

The main challenges for the proposed approach concern issues related to: landmarks, pose estimation, and learning method. These issues are briefly introduced in the following three subsections.

3.3.1 Landmarks

The type of landmark proposed in this thesis is a challenging topic due to the many related issues which need to be investigated. The type of texture represented in the patch as well as the patch size have big influence on system performance, so that those factors need to be carefully established when designing the **method for view acquisition and landmark extraction**. In particular, the landmark extraction needs to be designed in order to provide landmarks which are discriminant, i.e. high contrast and unambiguous. An attention selection mechanism tuned on desired type of texture and patch size, can represent an appropriate way of extracting landmarks when this is applied to images related to typical indoor environments. It then becomes important to design and verify the attention mechanism to understand which type and number of image features should be contained into extracted texture patches. The correspondence between image features and environment characteristics should also be analyzed. These topics are developed in chapter 5.1.

The method proposed for landmark recognition is vision based, that is, based on recognition of landmark appearances into the image-plane. The type and the quantity of image features contained in a texture patch then play an important role. But, which algorithm should be adopted for the **landmark match**? As seen in the literature an answer is typically represented by a compromise between resulting reliability and processing time, (see chapter 2). It is then very important to find out about the algorithm which best suits for the visual match through studies and experimentations. The proposed method must allow for recognition of landmarks previously acquired, from the current robot position. A main point is then represented by the way previously acquired landmarks should be processed to become comparable to a current landmark observation. The invariant characteristic of image features contained in a texture patch as well as the possibility for a **re-projection of landmark-appearances** previously acquired represent arguments to be carefully addressed. These topics are developed in sections 5.1.3, 4.1, and chapter 6.

Since the landmark representation is close to the image domain, the landmark appearance is likely to be not invariant. Consequently, the landmark shape, appearance, and discriminant characteristic, have to be analyzed in order to understand which is the best way of estimating **landmark positional information** This topic is developed in section 5.2.

Depending on the landmark invariance or on the possibility for a re-projection, the reliability of landmark recognition may be different for different workspace regions. A so called landmark "**observation area**", or "recognition field", needs then to be foreseen and experimented for typical landmarks in order to estimate which is the number of landmarks an environment should provide, which type of indoor environment represents a more suitable application context, and what reliability one should expect by the proposed approach. This topic is developed in section 4.1, and chapter 6.

In summary, the main challenges concerning landmarks are: landmark view acquisition and extraction, recognition and re-projection, pose computation, and observation-areas estimation.

3.3.2 Pose Estimation

The method for robot pose estimation is a challenging topic since much attention has to be paid on the resulting accuracy, robustness and error propagation. The performance of the proposed localization algorithm, based on triangulation and optimal triplet selection, needs to be tested in typical indoor setting, and the result analyzed in order to understand potential, limits and possible improvements.

A particular attention must be paid to setting camera parameters according to typical distances to landmarks and to the requirement of seeing three landmarks into the same image-plane. The results is a **required distribution of landmarks** in the workspace which represents an issue to be carefully analyzed. This topic is discussed in section 5.3.2.

The estimated robot pose is based on landmark appearance and their position and orientation in the workspace. Consequently, the fidelity of landmark appearance and the accuracy of landmark positional and shape information, become important factors to analyze and possibly quantify. The estimation of **error propagation in landmark visual and positional information** based on the uncertainty of robot positional information and matching algorithm, then represents important knowledge to gather and experiment with. This topic is developed in sections 5.3.

Once the effect of the main factors affecting the robot pose estimate has been understood, the issue of how to interpret and **best exploit landmark associated information** should be addressed. The result of such analysis should also affect the design of the method for optimal triplet selection. The method for optimal triplet selection represents a fundamental help towards providing accuracy and robustness to the localization system. It consequently is very important to understand which parameters among the many available, (fidelity, accuracy, relative position among landmarks, robot to landmark distance, etc.), should be weighted more than others for the triplet selection. In other words, to establish on which basis a triplet is elected as optimal. This topic is discussed and developed in chapter 4.

It is also relevant to examine the **limits to the main parameters** involved in robot pose computation. For example, if the uncertainty in previous robot pose estimation is above a certain limit, it is not possible to obtain an accurate estimate of landmark position by the proposed method. Analogously, if landmark orientation with respect to camera orientation is above a certain threshold, such a landmark should not be considered for robot estimation (the occurring perspective distortion would likely lead to an inaccurate, wrong, or misleading, landmark recognition). These topics are discussed and developed in section 5.3.

In summary, the main challenges concerning robot pose estimation are: fidelity, accuracy and distribution of landmarks; error propagation in landmark visual and positional information and main contributive factors to it; best exploitation of available information for landmark selection, (optimal landmark selection); and limits of main parameters involved in robot pose estimation.

3.3.3 Learning

The method proposed for the learning is a challenging topic because of the requirement of autonomous behavior and accurate estimation of landmark visual appearance and pose.

The **learning navigation strategy** needs to be designed as a procedure which can be applied to different situations arising at different execution time. For example, during system initialization when human assistance may be provided; during robot navigation when positional accuracy is very high, or in case this is very low. The learning navigation strategy also needs to be designed in order to make the robot follow paths which allow for: a low uncertainty arising from robot movements, (needed to extend system autonomy); a wide workspace observation, (needed to learn discriminant and spatially distributed landmarks); and acquisition of representative images of landmarks, (needed as reliable reference). In other words, a navigation strategy allowing for minimizing robot positional uncertainty and maximizing the number of learned landmarks. This will in fact affect precision and reliability of computed landmark poses. These topics are discussed and developed in section 5.4 and in chapter 6.

The way the developed learning strategy will affect landmark recognition during self-localization is another important factor to analyze and test. In particular, it is relevant to minimize the learning of erroneous information through verification ("**consistency checks**") at run-time, as well as the verification of estimated uncertainty associated with computed landmark poses. It is also relevant to examine the **consequences of uncertain learned information** related to landmark appearance and pose estimation in the recognition phase, as well as the consequence of an approximated landmark re-projection.

The development of the learning method has in conclusion to go through a careful analysis and testing. In summary, main challenges concerning the learning method are: learning navigation strategy, control of accuracy in the learning process, and estimation of the consequence of inaccuracy in learned information in the recognition process.

3.4 Research Development Plan

This section presents the proposed research development plan. The research activity has been planned through consecutive steps forward in order to efficiently study, develop, and experiment with, the many topics involved and the many issues presented in the previous section. In this way, each research step will be based on experiences and results gained at previous steps. The proposed development steps are: self-localization, automatic learning, automatic recognition.

1. **Self-Localization**

   This first step of research is for studying and developing the proposed self-localization method. In particular, implementation on a real mobile platform of the algorithms for pose estimation based on triangulation and optimal triplet selection.

   The world model is priori known at the time the robot perform the localization. In particular, landmark appearances will be learned by the system with the assistance of a human operator. The suggested technique is a manual training, (see section 1.5.4). Landmark positional information are then measured by hand and fed into the system. This is done to ease development of the system and serve as reference. In fact, this process will be automated in next development steps.

   Based on the provided information, the vision system will be required to identify three landmarks in the camera image-plane during navigation in order to calculate robot position and heading by a triangulation method. The starting point is a promising work in the state of the art providing a theoretical framework for optimal selection of landmarks triplets, (Madsen and Andersen, [97]), which could be adopted for landmark selection before triangulation takes place. However, this previous work does not demonstrate its application on a real robotic system. Consequently, the first step of the research in this thesis includes the implementation of the proposed theory in a real robotic system.

   The estimated robot poses will then be compared to ground-truth and to pose estimates achieved by using alternative methods from the state of the art.

   The lesson to learn is based on the analysis of system performance which will tell us about potentials and limits of the proposed method. The proposed approach for self-localization and the result of the experimentation have been presented in a publication, (Livatino and Madsen [93]), and they are thoroughly described in chapter 4.

2. **Automatic Learning**

   The second step of research is for developing, implementing, and studying on a real mobile platform, the proposed method for automatic learning of landmark reference images and poses.

   A method mainly based on panoramic view synthesis, attention selection, and stereo reconstruction, is proposed to be studied and developed during this research step. Note that in this step the system will not posses any prior knowledge about the workspace nor about the system's initial position.

   The learned landmark information will be compared to ground-truth to verify system accuracy. The uncertainty related to computed landmark pose is estimated based on the expected error on template matching algorithm and on a typical error affecting robot pose during localization, estimated at the previous research step.

The lesson to learn is based on the analysis of system performance in terms of "fidelity" of reference views and accuracy in estimated landmark positional information. This will tell us about the potential and the limits of the proposed learning method. The proposed approach to self-learning and the results of an early and a later experimentation have been presented in two publications, (Livatino and Madsen [92], [94]), and they are thoroughly described in chapter 5.

3. **Automatic Recognition**

   The third step of research is for studying, developing, and implementing on the robotic system, the proposed method for automatic recognition of self-learned landmarks. In particular, to design how to use the automatically learned landmark representative images and poses, in order to reliably match arbitrary landmark views arising during robot navigation.

   It is also necessary to verify that the acquired landmark positions and appearances are accurate enough to allow the system to predict during self-localization landmark appearances from arbitrary positions, and use such prediction to successfully recognize the landmarks in the incoming images. The system will consequently posses the knowledge of landmark representative images and positional information, learned at the previous research step.

   The main challenges in this case are represented by the errors associated with the learned landmarks as well as the errors affecting robot pose during its navigation. A method is then proposed based on "realistic" landmark virtual-views visualization and template matching by image-correlation .

   The usefulness of automatically acquired landmark views and estimated poses will be tested by analyzing the recognition performance and by comparing this to the results obtained at the first research step, when landmarks where manually and accurately provided to the system.

   The lesson to learn is based on the analysis of system performance which will tell us about the potential and the limits of the proposed method for automatic recognition, but also for automatic learning. The proposed approach for automatic recognition and the result of the experimentation have been presented in a publication, (Livatino and Madsen [94]), and they are thoroughly described in chapter 6.

Chapter 4

Robot Self-Localization

Localization is the process of finding the position and orientation of the robot (pose) relative to an external coordinate system. If other than estimating its pose the robot is required to autonomously perform the navigation task, then the robotic system also needs the capability of computing its pose automatically. This process is called *self localization.*

This chapter describes the first step of research suggested in the development plan of section 3.4, Self-Localization. The approach proposed in this thesis for autonomous navigation and in particular robot self-localization, combines the use of vision, landmarks and triangulation, (see chapter 3). In particular, the proposed solution for robot-pose estimation is based on: (1) automatic recognition of visual landmarks during navigation; (2) triangulation of recognized landmarks.

The two fundamental steps above mentioned are presented in sections 4.1 and 4.2, respectively. The proposed self-localization scheme, which integrates functionalities presented on both the two sections, is then described in section 4.3. Performed experiments are described and commented at the end of section 4.1, for what concerns landmark recognition, and in section 4.4 for what concerns the entire system.

4.1 Automatic Recognition of Visual Landmarks

In order to estimate robot pose by the triangulation method proposed in this thesis, (see section 4.2), at least three landmarks need to be located in the camera image-plane during robot navigation. Landmarks can be located in the camera image-plane by comparing them to previously acquired ones. This process is called *landmark recognition*.

In the application context of metric navigation and accurate pose estimation, the robot needs to accurately monitor its position and heading during navigation, which in turn means that the robot needs to estimate its pose continuously. A continuous pose estimation can not take place with the proposed self-localization method without a continuous recognition of landmarks. Landmark recognition during navigation is consequently highly demanded to be an *automatic process*. This would in fact allow for a continuous pose estimation, which in turn means that the robot can navigate an environment for an extensive period of time without the need for human assistance. The goal of proposed landmark recognition method consequently is to automatically locate landmarks in incoming images during navigation.

The proposed technique for landmark recognition requires that the landmarks we want to locate are "known" at the time they are observed in incoming images during navigation. Following the consideration made in subsection 1.5.4, landmark knowledge can be gained in different ways. It has been planned to first experiment with a manually learned landmark model in order to understand about performance of proposed triangulation method in this "favorable" acquisition context, and only then, in a next developing step, to experiment with automatically learned landmarks.

The goal of this section consequently is to introduce the basic techniques involved in the proposed automatic recognition method, whereas a later section (in chapter 6) will presents how the proposed method "evolves" in cases landmarks have been automatically learned by the method proposed in chapter 5. The two developing steps in landmark recognition also represent the way recognition method has been studied, developed, and implemented, by the author of this thesis.

This section starts with a brief analysis of main problems concerning with automatic landmark recognition, and it then follows with proposed solution model described in sections: landmark visual prediction, landmark matching, and observation area. Performed experiments are presented in the last subsection.

4.1.1 Problem Analysis

Landmark recognition is unfortunately challenged by many problems. They are mainly related to the fact that landmarks to be compared have been captured in different periods of time, and so likely they have been observed from different viewpoints. In addition, uncertainties arising from robot movements, algorithmic errors, the landmark acquisition process, etc., jeopardizes the possibility for a reliable landmark recognition. The main factors involved when observing landmarks from different viewpoints are thoroughly discussed in subsection 5.1.1. The method proposed for processing landmark information, (visual and positional), for their automatic recognition plays then an important role.

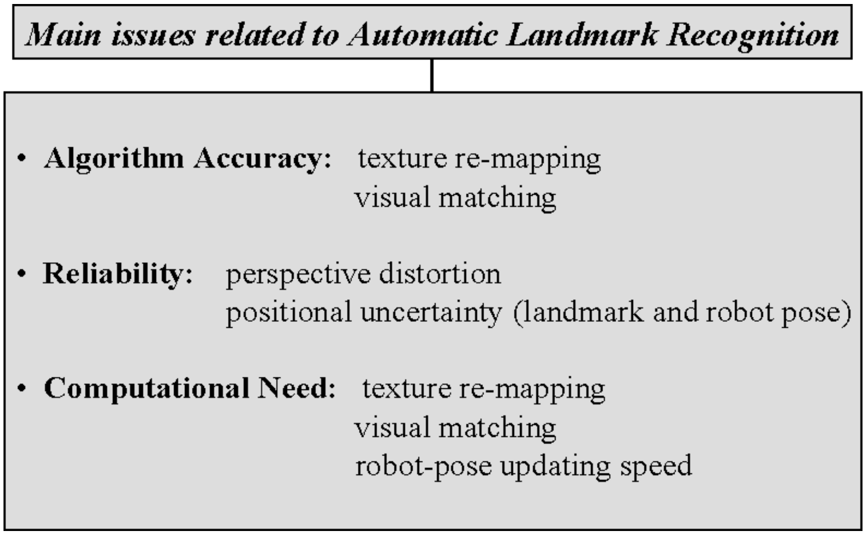
As discussed in subsection 3.1.3, it is proposed that landmark comparison takes place at appearance level, that is in our case, at "image level". During robot navigation, the landmark appearance in incoming images is called: *current view.* The landmark appearance acquired instead previously during a learning phase is called: *reference view.*

Unfortunately, being current and reference views taken in different moments, and so likely from different workspace locations, (the robot has moved), they might look different and can not be directly matched in incoming images. In particular, the discrepancy between previous and current viewpoint has likely become too large to allow a direct comparison.

In order to make the two landmark views comparable, it is proposed to compute the expected appearance of a landmark in a current view, which would so allow for a match in current image-plane. The expected appearance of a landmark in a current view and the landmark match in current image plane, represent then two important aspects to attention. In fact, the important issues are: (1) to understand how arising perspective distortions, and uncertainties in estimated landmark and robot poses, will affect a re-mapped landmark appearance; (2) how to best perform landmark matching in order to achieve a reliable recognition response. In other words, main questions to be answered are: which technique should be applied for landmark re-mapping and matching? How reliable is a performed match going to be?

Another critical aspect of the automatic recognition process concerns computation time. In particular, the complexity of the proposed re-mapping and matching algorithms, and how this would affect robot-pose updating speed. Note that the problems of estimating robot pose and pose updating-speed, are coupled. If positioning is more accurate then updating speed can be made less frequently. The algorithms used for re-mapping and matching then play an important role on system (localization) performance.

Figure 4.1 summarizes main issues related to automatic landmark recognition.



Main issues related to Automatic Landmark Recognition

- **Algorithm Accuracy:** texture re-mapping
  visual matching
- **Reliability:** perspective distortion
  positional uncertainty (landmark and robot pose)
- **Computational Need:** texture re-mapping
  visual matching
  robot-pose updating speed

Figure 4.1: The figure represents main issues related to automatic landmark recognition.

4.1.2 Landmark Visual Prediction

The available knowledge consists of an estimate of current robot pose, camera focal length in pixels, landmark position and orientation, and landmark reference views. Based on the available knowledge, it is possible to compute landmark visual appearance from current camera pose. In other words, it is possible to re-map reference views according to current viewpoint.

But which are the characteristics that a reference view should posses? and how can we acquire representative views?

A reference view should be a representative image of the visualized region. A representative view of an environment characteristic should clearly represent all texture details of the represented objects. This would in fact allow for a high-fidelity view re-projection, i.e. avoiding that this operation would generate incomplete or incorrect views, (due, for example, to new arose visible aspects). A representative view should also be taken as close as possible to the object of interest, so that a re-projected texture would always be the result of a compression, and not of an enlargement.

In case of a texture representing a planar surface, the most representative view is a frontal view. The ideal situation would then be for our case to capture occlusion-free frontal-views of landmarks, taken to a convenient distance. Figure 4.2 shows the ideal situation setup. Being landmark model obtained by a *manual training*, (see section 1.5.4), we know that landmarks approximately represent planar surface.

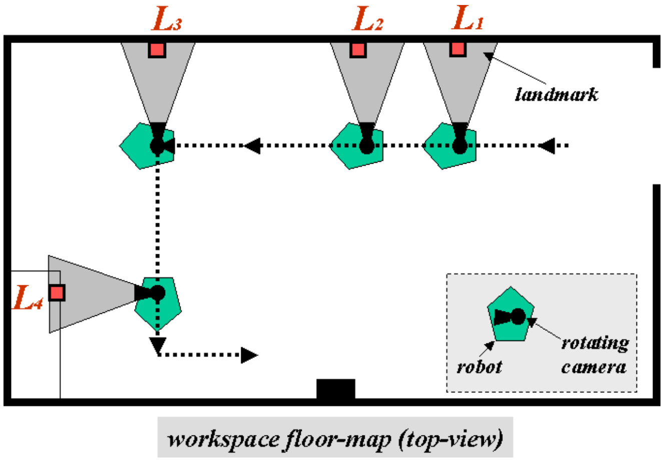


Figure 4.2: The figure represents the ideal situation setup for proposed *manual training*.

The view re-mapped from a reference-view view is called: *landmark visual prediction*. The figure 4.3 left and right hands shows examples of re-mapped views and the information required for the mapping. The landmark visual prediction can be estimated by a forward or inverse projection technique depending on computational-time constraints and required accuracy.

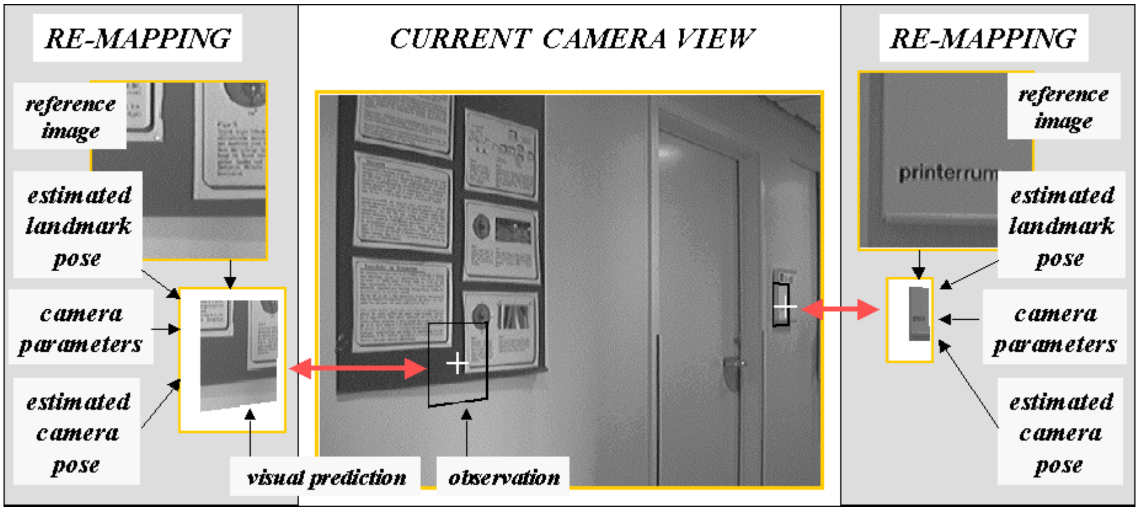


Figure 4.3: The figure represents a schematic representation of the landmark re-mapping process compared to observation.

Forward Mapping

In the case of a forward mapping of texture, that is, from reference view to current view, each point in the 2D reference view is projected to the 3D world and then to the 2D current view. Texture value in the current view is then set according to correspondent texture-value in the reference view.

The projections from 3D world to 2D image space and vice-versa are based on a *pin-hole* camera projection scheme. This is described in fig 4.4. This represents a common way to describe the transformation from space points in camera coordinates to image points, (Gonzalez and Wintz [58], Tsai [143], Störring [134]).

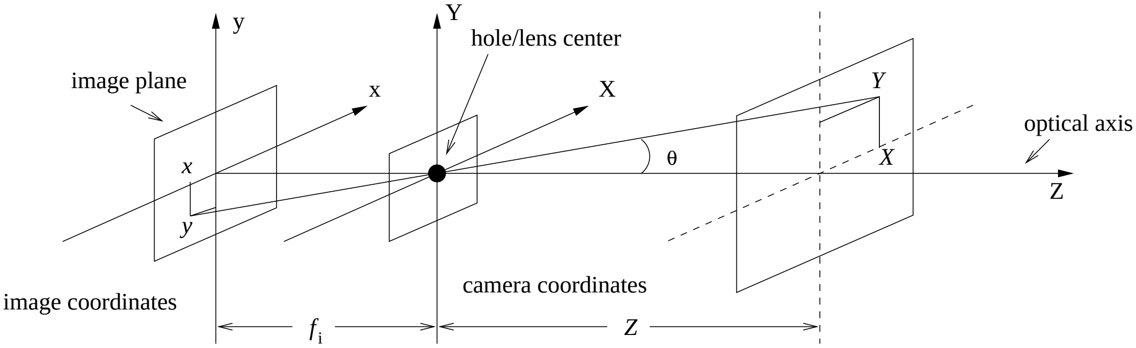


Figure 4.4: The figure represents image formation with the pin-hole camera.

The imaging element of this camera is an infinitesimal small hole. Only the light ray coming from a scene point at (X, Y, Z) which passes through this hole meets the image plane at

$(x, y, -f_i)$. Through this condition an image of an object is formed on the image plane. A straight line in the space is projected onto a straight line at the image plane. All object points on a ray through the pinhole are projected onto a single point in the image plane. The observed 3-D space is essentially reduced to 2-D surfaces.

The relationship between the 3-D camera and the 2-D image coordinates (x, y) is given by:

$$x = -\frac{f_i X}{Z} \qquad y = -\frac{f_i Y}{Z} \tag{4.1}$$

The two camera coordinates parallel to the image plane are scaled by the factor f_i/Z where f_i is called the focal length. The image coordinates (x, y) contain only ratios of camera coordinates, from which neither the distance nor the true size of an object can be inferred. The complexity of the re-projection algorithm for a forward mapping of texture is $O(N_{ref})$, where N_{ref} is the size of the landmark reference view.

Inverse Mapping

In the case of an inverse mapping, that is, from current view to reference view, each pixel in the 2D current view is back projected to the 3D world, and then to the 2D reference view, in order to find the correspondent value in the reference view. The texture value in the current view is then set according to its corresponding value in the reference view. The complexity of the re-projection algorithm for an inverse mapping of texture is $O(N_{cur})$, where N_{cur} is the size of the landmark in the current view.

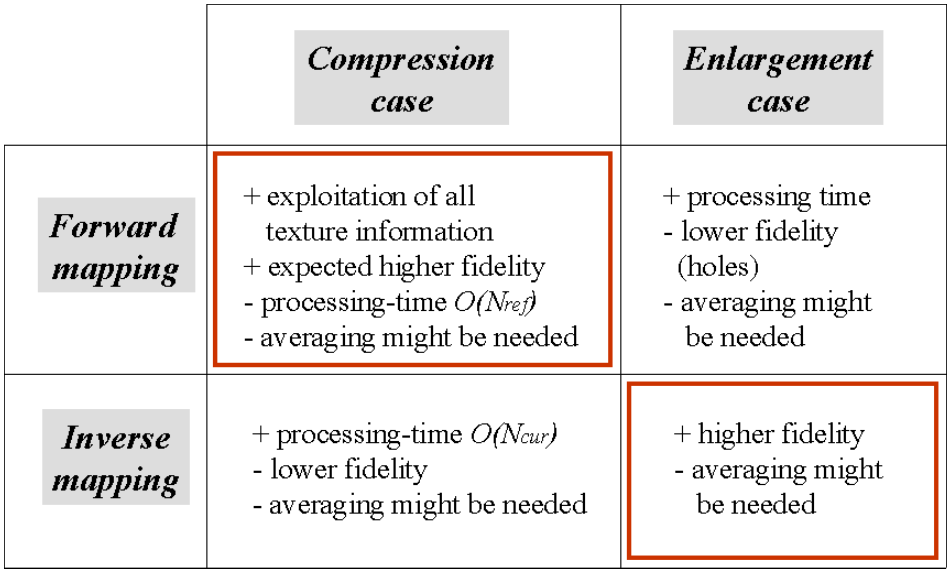
The resulting pixel correspondence between reference and current views is in theory the same both in a forward and an inverse mapping, however, in the latter the computation is done for each pixel in the current view rather than in the reference view. This influences processing time as well as texture fidelity.

Comparison

The disadvantage of a forward mapping over an inverse mapping, in the compression case, (i.e. when current view is smaller than the reference view), is that several pixels in the reference view would be mapped to the same pixel in the current view, and this could result in a "waste" of computational time. The map will in fact cost the same computational amount regardless of landmark size in current view. In addition, some sort of averaging is required to set a texture value for those cells related to several pixels in the reference view. An inverse mapping would instead reduce the computational time in the compression case. However, the texture fidelity could be lower with an inverse mapping.

The disadvantage of a forward mapping in the enlargement case, (i.e. when the current view is wider than the reference view), is that texture values in some image-regions of the current view could not be mapped, (due to the "pixel-size" approximation), and this leads to "holes" in the resulting view which would require some extra processing to fill the "holes" in. Figure 4.6 right-hand shows an example of computed landmark visual predictions in the enlargement case, emphasizing undesired mapping problems.

In summary, the advantage of a forward mapping over an inverse mapping is in the compression case in terms of accuracy of the visual prediction, since texture fidelity might be higher



	Compression case	*Enlargement case*
Forward mapping	+ exploitation of all texture information + expected higher fidelity - processing-time *O(Nref)* - averaging might be needed	+ processing time - lower fidelity (holes) - averaging might be needed
Inverse mapping	+ processing-time *O(Ncur)* - lower fidelity - averaging might be needed	+ higher fidelity - averaging might be needed

Figure 4.5: The summarizes main advantages (+) and disadvantages (-) of a forward and inverse mapping. The cells emphasized in red indicate chosen cases.

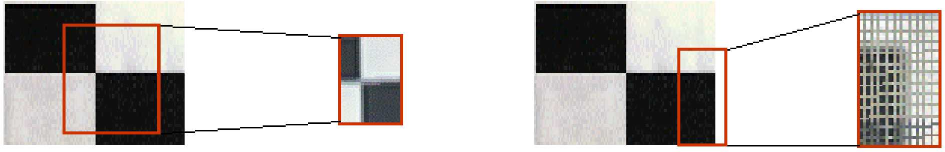


Figure 4.6: The figure shows example of computed landmark visual predictions, both in case of compression and enlargement, emphasizing undesired mapping problems.

due to the exploitation of all available texture information. The disadvantage of the forward mapping is in the compression case, less computational efficiency, and in the enlargement case, the generation of "holes" in the resulting view. Analogously, it is possible to state that the advantage of an inverse mapping over the forward mapping is consequently in the compression case computational efficiency and in the enlargement case, a potentially higher texture fidelity. The table 4.5 summarizes main advantages and disadvantages of a forward and inverse texture projection.

Since our aim is a high texture fidelity, it is proposed to adopt a forward mapping in the compression case and an inverse mapping in the enlargement case. Experiments showed that the proposed re-mapping algorithms performed well in the considered application context. It seemed that in the compression case a forward mapping provided better results than an inverse mapping, whereas in the enlargement case an inverse mapping was to be preferred. That is, as expected. However, when the robot positional error was small, the difference between forward and inverse mapping was difficult to appreciate for discriminant texture patches and the resulting visual predictions showed high fidelity.

4.1.3 Landmark Matching

It is proposed to use template matching techniques in order to recognize landmarks, and so locate them in current camera views. Template matching can provide a very accurate information but we need a template. A template can be represented in our case by the landmark visual prediction.

There are many algorithms proposed in the literature concerning image matching. The comparison between two images can either be feature or correlation based. The correlation based technique is preferred due to its characteristic of a reliable texture recognition.

The algorithm used to perform the matching of different views of landmarks, plays an important role towards a reliable recognition, (as well as landmark dimension and type of texture). Among different correlation techniques proposed in the literature for match of images, it has been preferred to use the *normalized cross correlation.* This due to its characteristic of reliable texture recognition.

The normalized cross correlation has in fact demonstrated to be very robust towards many type of distortions, (Aschwanden and Guggenbuhl [5], Martin and Crowley [100]). In particular: distortions due to "moderate" illumination changes ("iris sequences"), zero mean Gaussian noise, noise arising from occlusions ("salty noise"), noise induced by varying the distance between sensor and object ("zoom noise"). The normalized cross-correlation is then proposed since:

1. texture patches representing landmarks are supposed to contain a rich and discriminant information.
2. texture patches with the proposed landmark-size range, can reliably be matched in images by using the normalized cross-correlation technique.

The formula related to normalized cross-correlation technique is shown below:

$$NCC(x,y) = \frac{\sum_{v=0}^{Vmax} \sum_{u=0}^{Umax} R(u,v) \cdot S(x+u, y+v)}{\sqrt{\sum_{v=0}^{Vmax} \sum_{u=0}^{Umax} R^2(u,v) \cdot \sum_{v=0}^{Vmax} \sum_{u=0}^{Umax} S^2(x+u, y+v)}} \quad (4.2)$$

The main parameters of this formula, which require to be carefully set, are:

$R(u,v)$ That is, the landmark visual template or image reference pattern.

$S(x,y)$ That is, the search window or region of interest. S represents the image-region examined when looking for a match between landmark visual template and the image.

$R^2(u.v)$ That is, the energy of the landmark visual template. It is calculated as in Martin and Crowley [100].

$S^2(x,y)$ That is, the energy of the search windows. It is calculated as in Martin and Crowley [100].

The size of a search window, S, has to be carefully set. This in fact represents a compromise between computational time and landmark recognition performance. In particular, a larger search window means a higher probability to reliably locate the landmark visual template in the image plane, while a smaller window means a faster computation but less trustworthy.

The computational demand of a matching increases with the size of the search window . A larger search window also increases the possibility of "catching" something else than the object of interest. Therefore, the size of the search window is matter, and it should not be larger than the area in which the object is "expected" to be in. The size of the search window in our experiments has been set according to expected error in robot position, which results in an uncertain image region around the expected location of the landmark in the image-plane.

A suitable size for a search window has been found through experiments. In particular, by considering typical distances to landmarks and expected error a search window set to a width double of the visual template width and a height 1.5 times of the visual template height, was considered sufficient. A typical landmark texture patch size in our experiment had size ranging from 28x28 to 36x36 through an extensive number of runs, (see section Experimentation). The search window size was consequently ranging from 56x42 to 72x54.

The existence of low cost correlation hardware and optimized algorithms, (see for example the frequency based approach proposed by Yeshurun and Schwartz [149]), makes cross-correlation a very attractive option. The drawback of expensive computation can so be reduced to around one or higher order of magnitude.

Despite the above mentioned advantages, the use of normalized cross-correlation is challenged by scale errors and it is computational expensive. In particular, the latter points out the issue of how to limit the size of a search window.

A prediction of landmark location into the current view can be estimated priori to the match. This can be done based on the known information, (i.e. position and orientation of landmark surface, camera focal length in pixels, and an estimate of robot current pose). Naturally, there may well be a discrepancy between the predicted and observed image location of the landmark, so that the expected landmark location does not match with the observed location. This is due to errors from robot movements and uncertainties in estimated landmark and robot poses.

Nevertheless, the landmark predicted location usually represents a close approximation of the observed location, so that it is sufficient to run the normalized cross-correlation algorithm in a circumscribed area around the predicted location. This avoids searching for a match through the entire image-plane. Figure 4.7 represents main "actors" in the matching process.

Experiments showed that normalized cross-correlation performs well on most of the cases with discriminant texture patches, since the system was able to reliably recognize valid matches based on the analysis of the correlation coefficient. More detail are provided in the experimentation section.

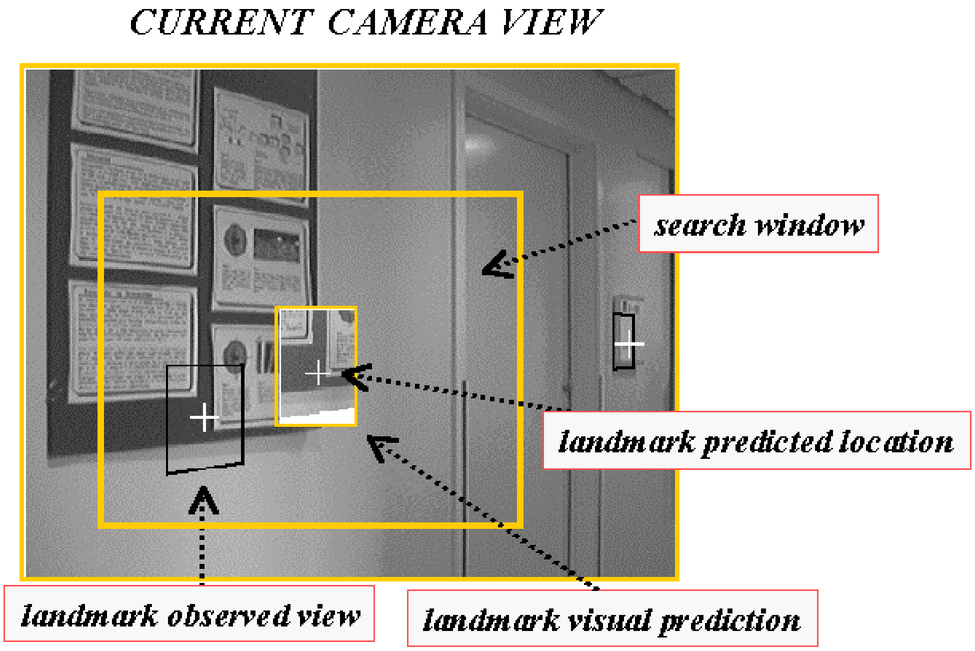


Figure 4.7: The figure represents a schematic representation of the landmark matching process.

4.1.4 Observation Area

An interesting issue when experimenting with proposed landmark matching technique was to observe that over a certain distance and angle to the observed landmark surface, no matches were allowed any longer. The following questions consequently arise: how far from the original acquisition viewpoint, (i.e. related to the reference view), can the robot run in order to still be able to recognize the landmark? which are main parameter involved in the automatic recognition?

The possibility during navigation for a successful and reliable match between landmark visual prediction, (i.e. landmark predicted appearance), and landmark observation, (i.e. landmark current view), depends on many factors, as it will be described in section 5.1.1. The focus of this subsection is on studying how ”distant” from the acquisition viewpoint can the current viewpoint lie in order to allow for a reliable match, hence, to introduce the concept of *observation area*.

When a landmark reference view is mapped to a new viewpoint, the size of the resulting image will likely be different from the original one. The size of the landmark visual prediction represents then an important parameter to take into account, (other than texture structure). In fact, if size of the visual prediction is below a certain lower-bound then there is a risk that generated textures do not provide enough discriminant information for a reliable match. This would then be due to the small resulting size and consequent loss of information involved by the texture compression.

If size of the visual prediction, on the other hand, exceeds a certain upper-bound, the risk would than be that generated textures become very approximated by the system trying to extrapolate "too much information" from the available patterns.

The so-called *perspective projection angle*, i.e. the angle between camera image-plane and landmark surface, (see $pp - angle$ in figure 4.8), also represents a critical factor since it may involve compression and enlargement, and perspective distortion may arise as well. Consequently, the angle value should be contained into a certain range.

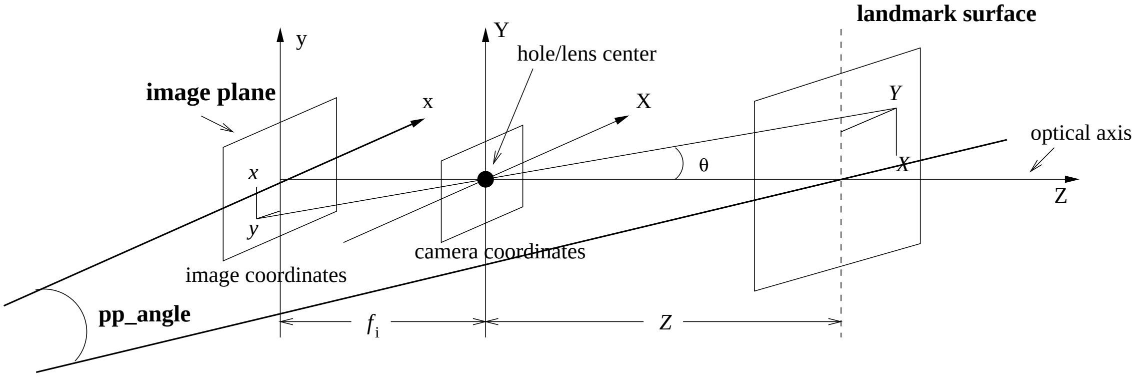


Figure 4.8: The figure illustrates the *perspective projection angle*.

It is believed important to test matching performance of the proposed recognition method according to different observation viewpoints, hence, to determine threshold values to upper and lower bound size of the visual prediction as well as to set the allowable range of projection angles. The result is a workspace area where a landmark is expected to have a high probability of being recognized without mismatches. The resulting workspace area is called *observation-area*. Figure 4.9 right-hand shows an example of observation-area expected for the landmark represented in figure left-hand. What defines the observation area shape is the set of robot positions which registered successful matches between visual prediction and observation. A robot position is then inside the observation area in case of a successful match.

The issue of "observation area" has also been discussed in some literature works. For example, Balkenius [9] refers to it as "view-fields", whereas Davison and Murray [41] tell about a portion of the workspace where a landmark can be identified from every position.

Experiments showed that recognition performance is different for different landmark textures. In particular, a better performance is related to textures containing large uniform areas, which shows the importance of considering local symmetries during landmark extraction. Tests with the specialized landmark (e.g. patch a in figure 4.13), showed the advantage of having a texture which is symmetric and which allows for a scalable template.

In conclusion, the knowledge of the observation area is very useful to decrease the probability of mismatches and missed recognitions, and so reducing noise in estimated robot poses. In particular, by knowing current camera pose and observation areas of previously acquired landmarks, the system is able to test the current observation conditions, and consequently do not consider a landmark for localization purposes when the current camera position falls outside the landmark observation area.

The knowledge of observation area could also be exploited to drive robot navigation by

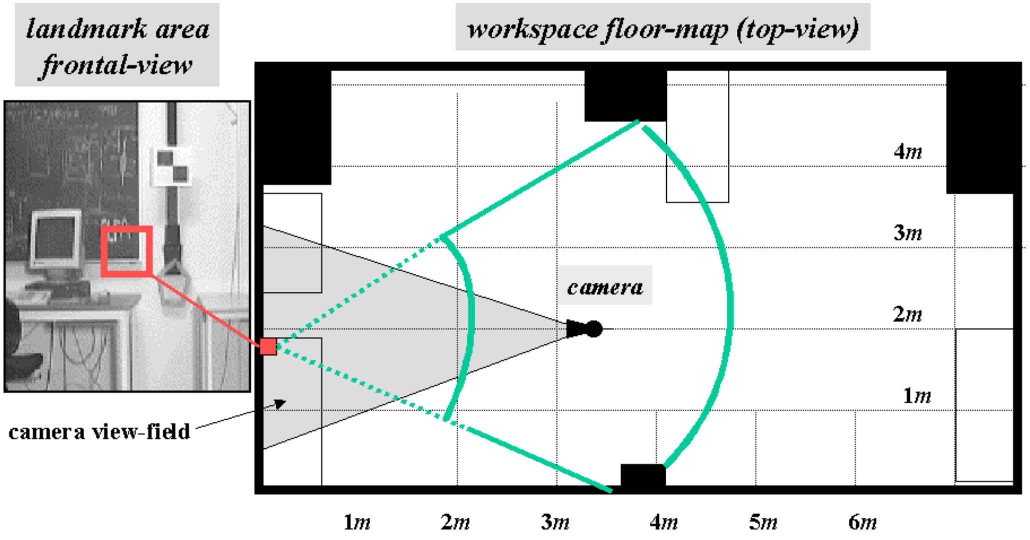


Figure 4.9: The figure right-hand shows an example of observation-area expected for the landmark represented in figure left-hand.

designing navigation paths which would improve recognition performance, and so localization accuracy.

4.1.5 Experimentation

This subsection describes the experimentation phase concerning automatic landmark recognition during navigation. The goal was to receive a feedback about performance of the proposed method for different types of landmarks in a realistic setting. In particular, to understand how different texture structures for the landmarks would affect recognition performance.

The algorithms were consequently implemented and tested on a real mobile robot system. The hardware used for the experiments was a Pentium 150MHz mounted on mobile platform (Robuter by Robosoft) equipped with a vision system consisting of a monocular CCD color camera sitting on a pan-tilt unit placed at the top of the robot, (so able to rotate over 360 degrees). Figure 4.10 shows the robotic system and the pan-tilt unit.

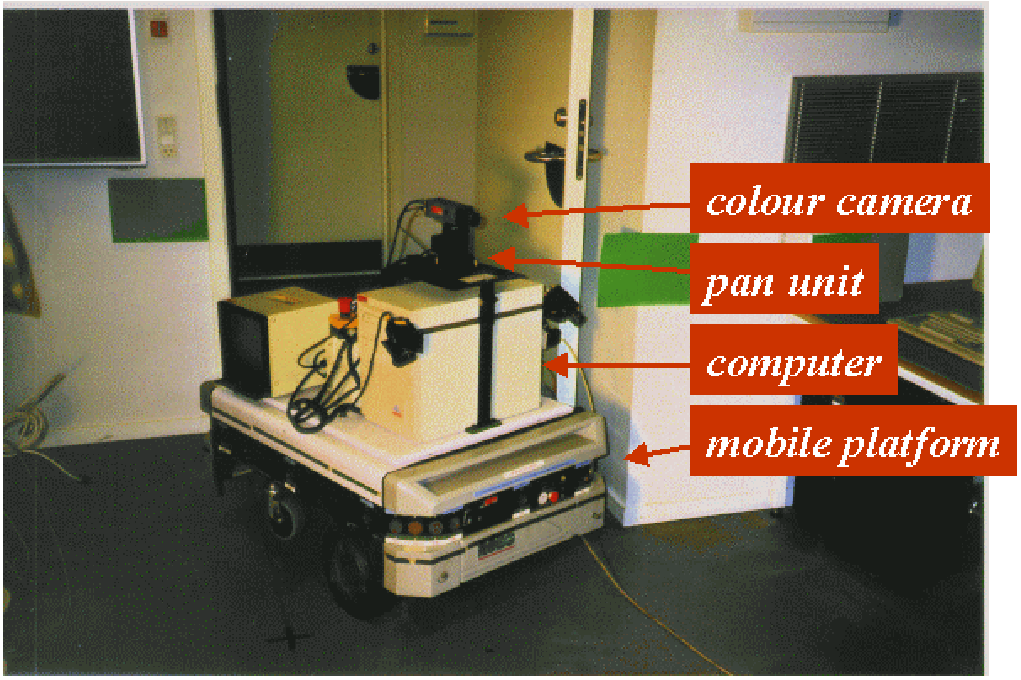


Figure 4.10: The robotic system used for the experiments.

The color camera resolution was set to 512 x 512 pixels. The internal camera parameters were estimated by observing the specially designed calibration object shown in figure 4.11 left-hand, and then by running an algorithm to recognize the object, (Bjornstrup [13]), based on the Tsai method, (Tsai [143]).

The system performance were tested in two different environments. They were the old and new laboratory rooms at the Laboratory of Computer Vision and Media Technology, Aalborg University. In particular, the old laboratory is represented by a two rooms $4m$ x $5m$ each with plenty of furniture and connected by a large doorway, figure 4.12 top-row shows the old laboratory floor-map; and the new laboratory is represented by a single $10m$ x $6m$ room, containing few furniture and few poster-pictures. Figure 4.12 bottom-row shows the new laboratory floor-map.

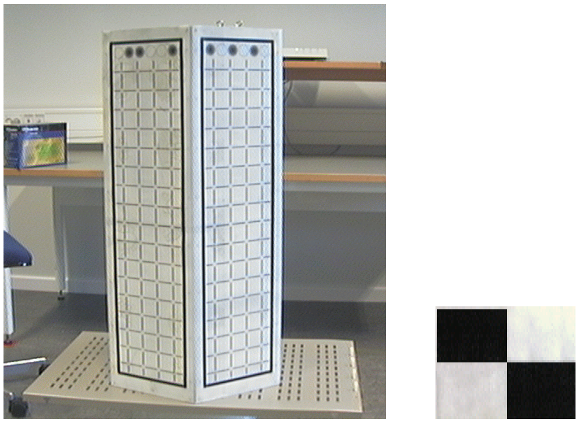


Figure 4.11: Figure left-hand shows a specially designed calibration tool used used for camera calibration. Figure right-hand shows the specialized landmark introduced in the environment to test system performance.

The different landmark textures tested in our runs are represented in figure 4.13. The figure represents texture patches related to both the specialized landmark and naturally occurring objects. In particular:

- The specialized landmark (patch a) was introduced in order to test performance for an ideal texture structure. In fact, other than being a high-contrasted texture showing clear edges, corners and symmetries, this texture patch is tolerant to scale errors, which represents a potential weakness of the proposed matching algorithm.

- The texture patch b, i.e. a bottom corner of an hanging black-board, represents a natural occurring object, but still with a high-contrasted texture and wide uniform symmetric regions. An interesting characteristic of the represented object is that landmark surface is not totally planar but a part of it, lies about 5 cm disclosed from the other part.

- More distinctive textures as well as naturally occurring, are represented by patches d and e. The patches contains uniform symmetric regions related to a door-handle and a key-lock. The landmarks surfaces contain a relief due the door handle. The handle sticks about 10 cm out of the door and it generates light reflections as well.

- The patches labeled f and g represent less favorable textures since the contrasted regions in the patches do not represent uniform symmetric patterns. These patches are consequently unsuitable for the proposed system. However, they would challenge the recognition method, which thus was the reason they were introduced in the test.

- The patch h represents part of the room wall and of a hanging poster. The patch size is the smallest (25x25 pixels) and contains some small but uniformly textured regions.

- The patch i representing part of the computer monitor is an almost flat surfaces with clear edges and wide uniform regions, however, it represents a challenge for the recognition method since the surfaces surrounding the landmark are not on the same plane.

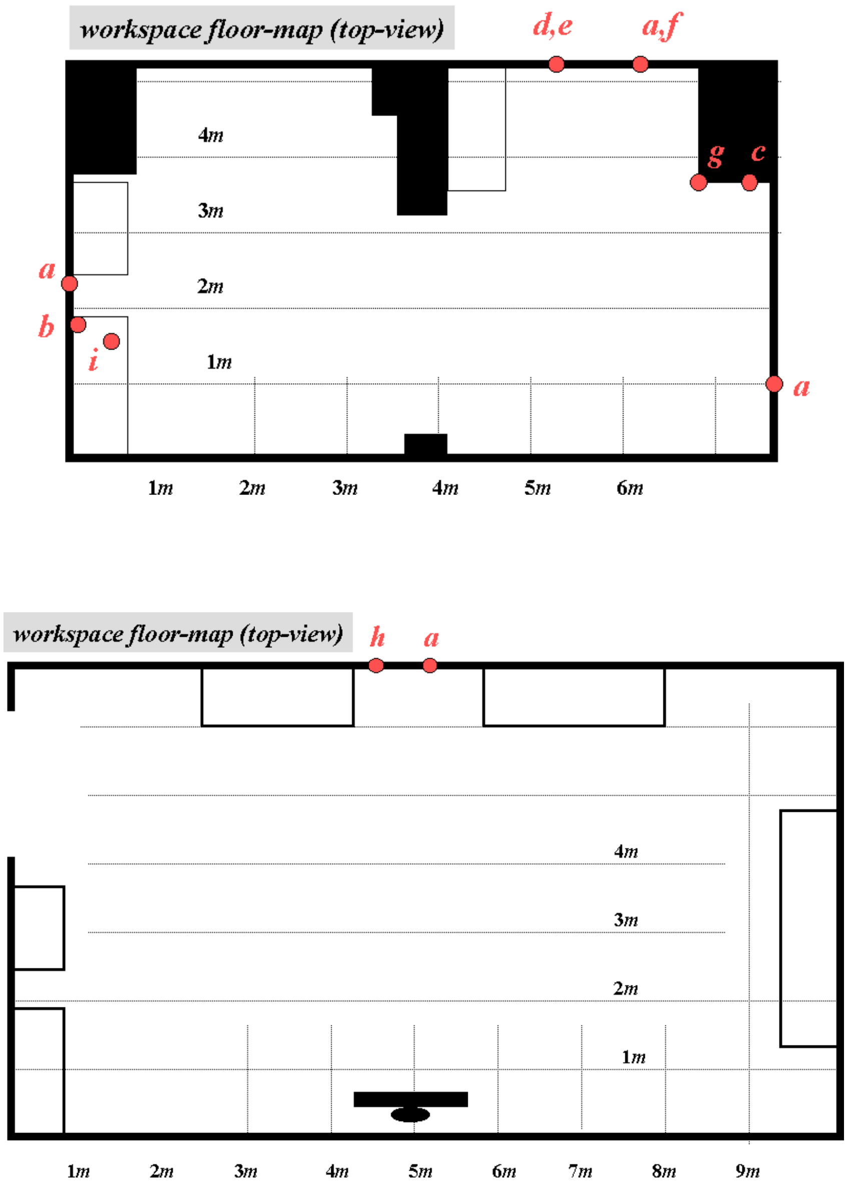

workspace floor-map (top-view)
d,e
a,f
4m
3m
2m
1m
g
c
a
b
i
a
1m
2m
3m
4m
5m
6m
workspace floor-map (top-view)
h
a
4m
3m
2m
1m
1m
2m
3m
4m
5m
6m
7m
8m
9m

Figure 4.12: The figures top-row show the workspaces used in the experiments, (respectively, the old and the new laboratory). The red labeled red dots represent the position of landmarks of figure 4.13.

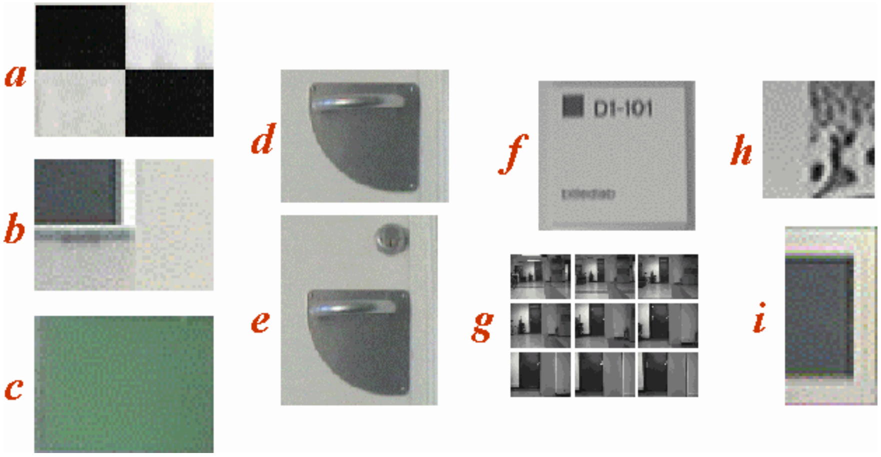


Figure 4.13: The figure shows the considered landmark reference-views.

- The uniform green-colored texture patch, labeled c, is very ambiguous and so unlikely to be recognized by the system. It was only introduced to verify arising errors in the size of re-mapped textures.

The information associated to ”previously” acquired landmarks, (pose and reference views), was provided to the system prior to the runs. This information was precisely estimated to avoid that uncertainties in these entities would have affected the final result. The positional information related to the landmarks were measured by hand with a tape meter. It was estimated that the real pose can be measured by hand with an accuracy of approximately 1 cm.

The landmark reference-views were acquired by manually driving the robot so that acquired views were represented by occlusion-free frontal views, taken to a convenient distance (around 3 meters). This is the case of acquisition by a *manual training*, and represents in this case an ideal situation as it was discussed in section 4.1.2 and described in figure 4.16 top-row. In particular, figure 4.16 shows floor-map of the workspaces containing landmark poses and acquisition trajectory for the manual training. Note that different landmark textures could be related to the same workspace area due to the fact that texture patches a, c and g where printed in a paper sheet so that it could be easily placed and removed depending on the experiment run.

The robot stopped for each landmark-view acquisition and correspondent robot pose was measured by hand with a tape meter, (average error around 3 cm.). We will refer to this situation as static pose-estimation. Note that robot pose was estimated with less accuracy than landmark pose because the complexity of the hardware devices made difficult to deduct the robot center. The pose of camera optical center (on-board the robot) could then be inferred knowing relative positions of the robot devices.

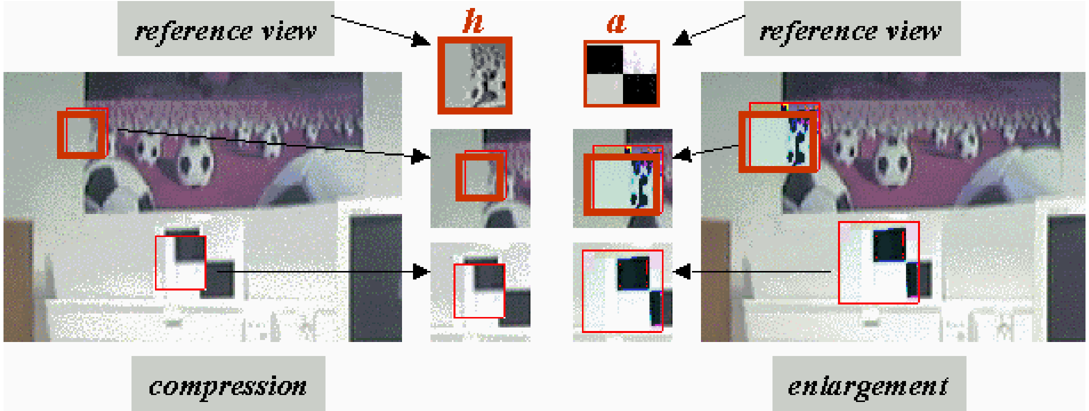


Figure 4.14: The figure shows examples of how a visual prediction related to same mapping technique (inverse mapping) could be effected by different errors depending on the current case (compression on the left-hand side and enlargement on the right-hand side). The visual predictions are superimposed to the observations.

Landmark Visual Prediction and Matching

The proposed algorithms for estimation of the landmark visual prediction from current camera viewpoint were implemented both for a forward and for an inverse mapping of landmark reference views.

Depending on type of texture and applied algorithm some artifacts were generated in the resulting views. These artifacts were mainly arising from errors in positional information associated to landmarks, current estimate of robot pose, and from algorithmic errors (leading to approximations).

Figures 4.14 and 4.15 show examples of computed landmark visual predictions and arising artifacts. In particular, the figure 4.14 shows examples of how a visual prediction related to same mapping technique (inverse mapping) could be effected by different errors depending on the current case (compression or enlargement); The visual predictions are superimposed to the observations so that the discrepancy between predicted and observed landmark locations can be noticed. Other evidences of inaccuracies are shown in the degeneration of the texture structure.

In the performed experiments the mapping result was apparently better for a compression by a forward mapping and for an enlargement by an inverse mapping. For example, figure 4.14 shows how the effect of mapping inaccuracy can easily be noted in the enlargement case and not in the compression case. Nevertheless, in cases of small positional errors and limited translation ranges, there were not significant differences in re-mapped textures of patches a, b, c, d, e, h and i. This happened in most of the cases in this first experimentation session also due to the small range of required transformation.

Despite degeneration of texture structure could sometime be observed, (see figure 4.15), it was not always immediate to understand its effect in the recognition performance. It was so proposed to analyze matching results, in particular correlation coefficient and estimated landmark location, in order to understand the effect of re-projection errors and so the reliability

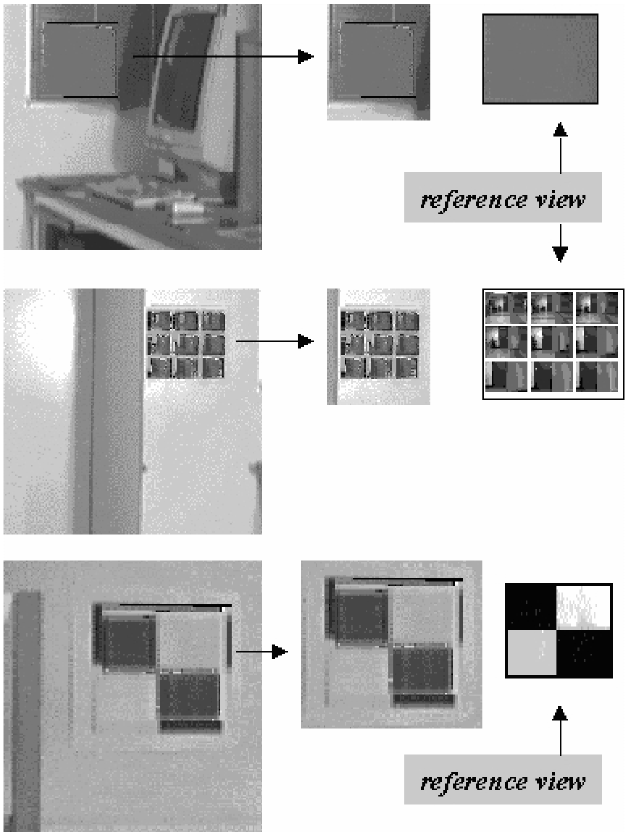


Figure 4.15: The figure shows examples of how a visual prediction could be effected by inaccuracies related to the mapping technique and the considered case. The top row of the figure represents the compression case for a forward mapping, and the middle and bottom rows represents respectively the compression and the enlargement case for an inverse mapping. The visual predictions are superimposed to the observations.

of proposed recognition method.

The landmarks were observed and matched to their visual prediction while the robot was let to navigate the environment. In particular, the robot was let to run along a rectilinear and relatively short trajectories, (around 5 meters). Robot initial pose was measured by the static pose estimation technique described in section 4.4.

Robot pose during navigation (dynamic pose estimation) was updated based on odometric readings. In this case the positional uncertainty was estimated taking in account expected uncertainties in: robot initial pose, odometric readings and the navigation trajectory. Large search windows were adopted in this case to cope with odometric inaccuracies. Nevertheless, because of the rectilinear trajectories and the relatively short distances traveled by the robot, the errors in odometric readings appeared to be limited to few centimeters. The knowledge of robot pose during navigation was required to turn the pan unit in order to make the camera looking at the landmark we wanted to recognize, as well as to re-map textures to the visual prediction.

Around 20 runs were performed and different landmark were tested at each run. Figure 4.16 bottom-row shows the navigation trajectories followed during the landmark matching together with camera poses at the moment landmarks were observed.

Experiments showed that normalized cross-correlation performs well on most of the cases and the system is able to reliably recognize valid matches based on the analysis of the correlation coefficient. In particular, the system were always able to successfully recognize the specialized landmark, (patch a in figure 4.13), if this was the only one of its type contained in the search window. The correlation coefficient was very high for this match, with values ranging between 0.7 and 0.97. A different range width, (between 0.53 and 0.89), was registered in case of landmarks representing a door-handle, (patches d and e in figure 4.13).

An interesting issue was to observe that the specialized landmark could be confused to other specialized landmarks in the room if size of the search window would allow to observe more than one specialized landmark. On the contrary, natural landmarks as the patches d and e, were search window independent since they are unique in a large area of the workspace.

A "poor" response was instead registered in case of landmarks represented by the patches f and g in figure 4.13. In particular, the patch f was successfully matched in only about 25% of the cases, (with a correlation coefficient around 0.65) and mismatches arose as well, (in this case the correlation coefficient was higher than 0.65). As for patch g, it was successfully matched in about 50% of the cases (with a correlation coefficient around 0.65), and never mismatched. Our guess was that patch g performed better than patch f because its texture structure shows some kind of regularity and contains sharper edges.

In case of patch h the matching response was successful in most of the cases. In fact, the smaller landmark template size did not seem to challenge the match, it was probably due to observation point next to frontal. In case of patch i the matching response was successful in most of the matches. However, performance are expected to be very different when observation point was less convenient, (see experimentation with observation areas).

The results of the experimentation are summarized in figure 4.17.

The landmark location in the image-plane was accurately estimated by the matching process (based on normalized cross correlation). The measured error in our runs was manually measured using the software tools *Imgview* and *Xv* provided in SGI computers. The average

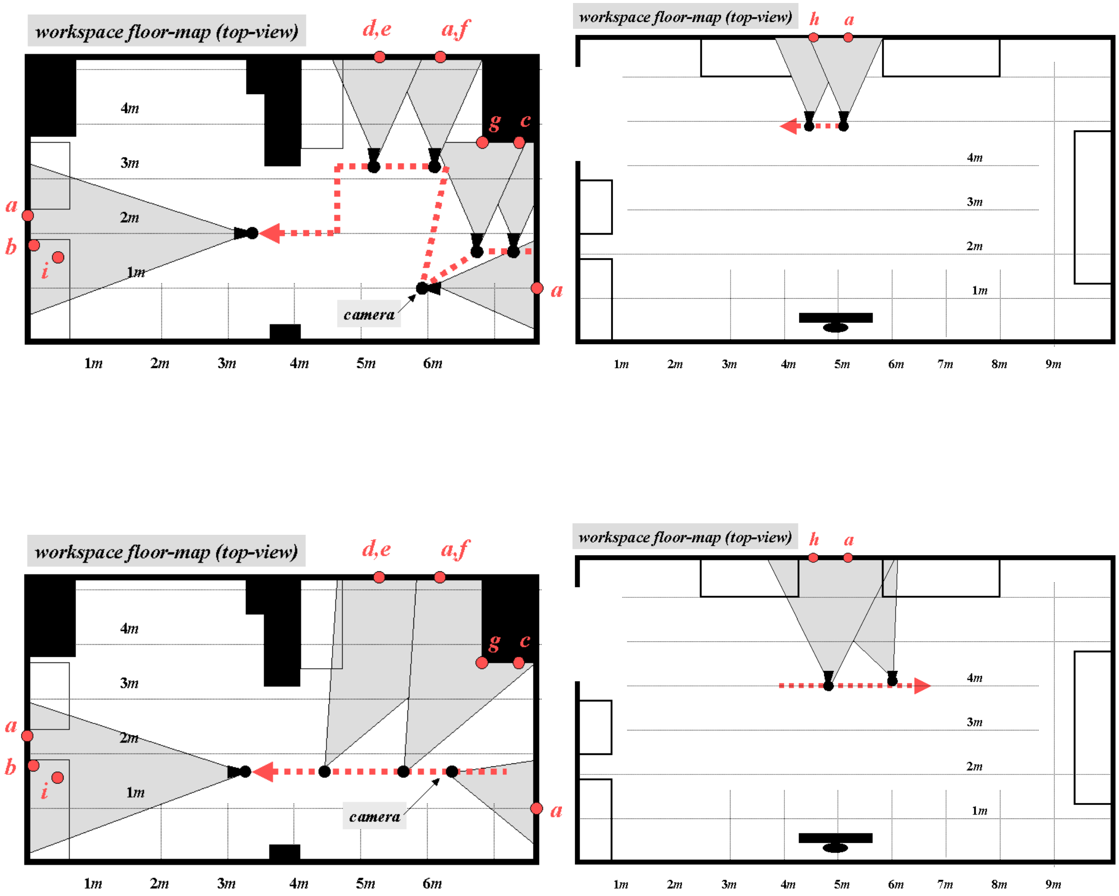
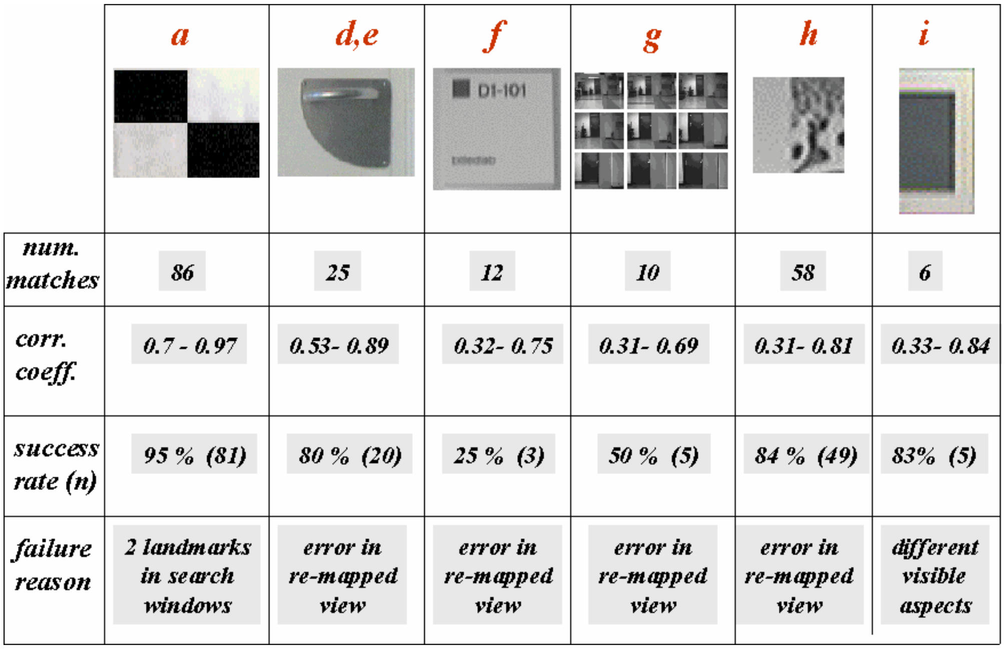


Figure 4.16: The figures show floor-map of the experimentation workspaces (old laboratory on the left-hand side), containing landmark poses. The figures top-row show the acquisition trajectories followed during the *manual training* together with camera poses at the moment landmarks were acquired. The figures bottom-row show the navigation trajectories followed during the landmark matching together with camera poses at the moment landmarks were observed.



	a	d,e	f	g	h	i
num. matches	86	25	12	10	58	6
corr. coeff.	0.7 - 0.97	0.53- 0.89	0.32- 0.75	0.31- 0.69	0.31- 0.81	0.33- 0.84
success rate (n)	95 % (81)	80 % (20)	25 % (3)	50 % (5)	84 % (49)	83% (5)
failure reason	2 landmarks in search windows	error in re-mapped view	error in re-mapped view	error in re-mapped view	error in re-mapped view	different visible aspects

Figure 4.17: The table summarizes results of the experimentation including texture patches, number of matches, correlation coefficients, success rate, match failure main reason.

error for all successful matches was roughly estimated around 2 pixel, which confirmed the accuracy of the normalized cross correlation when correlation coefficient is above the value of 0.55.

The time required by the matching operation was around 3 seconds for a template of 32x32 pixels in a search window set to 96x56 pixels. The algorithm ran a not optimized implementation of normalized cross-correlation on the available hardware (described at the beginning of this subsection). It was possible for the specialized landmark to successfully use a smaller template of 13x13 pixels and consequently speed up the matching process. This thanks to the texture characteristic of being symmetric and scalable.

The overall result were very encouraging and as expected. The landmarks considered discriminant and so suitable to our method, were in fact successfully and robustly recognized with a correlation coefficient far higher than the set threshold of 0.55. The experiments results also demonstrated that the proposed re-mapping algorithms performs well.

Observation Areas

The goal of this first experimentation concerning observation areas, was to understand about the shape of an observation area and how its size would be affected by different textures. An absolute estimation of observation area size was instead left to a later experimentation under a more realistic pose estimation setup.

Three landmarks were used for this test. The related texture patches are shown in figure 4.18 left-hand side and they correspond to patches a, b, and i in figure 4.13. The correspondent landmark poses in the workspace are shown in figure 4.18 right-hand side. The three landmarks were chosen because they represented different types of landmarks, and at the same time, they could be observed from the same camera pose. The latter would in fact simplify the experimentation, as well as the understanding of achieved results, (observation areas could easily be compared).

The specialized landmark was expected to allow for recognition in wide area because of its high-contrasted texture and its tolerance to scale errors. The landmark representing a blackboard corner was also expected to allow recognition in a wide area (though, smaller). This landmark represents a more natural landmark, and it contains a relief of about 5 cm which is expected may have influence on recognition performance, especially in cases of lateral observations. The landmark representing part of a computer-monitor is expected to be the most difficult to recognize since, despite it represents flat surfaces with clear edges and wide uniform regions, the surroundings surfaces are not on the same plane. The latter means that landmark visible aspects will change when the landmark is observed from position different than frontal, and the recognition performance would consequently be challenged by this fact.

The reference views associated to the three landmarks were taken from the camera pose, (showed in figures 4.18 left-hand). The distance between the camera and each landmark was around 3.5 meters. The acquired landmark reference-views had a template size of 32x32 pixels in cases of the specialized landmark and the one representing a blackboard-corner, and 22x40 pixels for the landmark representing a computer-monitor.

In order to test how the generated perspective distortion might have affected landmark recognition, landmarks were observed from selected viewpoints in a grid. The robot pose was estimated for each selected viewpoint by the static pose-estimation described in 4.4. The chosen viewpoints in the grid are shown in figure 4.18 as little circles.

Perspective transformations applied to reference view textures may introduce distortions depending on the type of texture, re-mapping algorithm, distance between acquisition pose and current pose and perspective projection angle. This was expected to bring the system to a point where matching was not possible any longer.

For each selected grid-viewpoint, the observed landmark texture (the current observation) was compared to the correspondent predicted texture, (the visual prediction), and we noted the grid positions where the match was successful or failed.

The conclusion of the experimentation is an allowable range for landmark texture distortion. In particular, the landmark should not be compressed to more than about a half of its original size for any of its dimensions, (width and height). Below this value, the landmark texture seems in fact to lose most of its texture characteristics so that it likely leads to mismatches or no match. In world coordinates a compression to a half of its original size for any of its dimensions, (width and height) means to double the distance between camera optical center and landmark center. Analogously, the landmark should not be enlarged to more than double of its original size since the effect of the average filter could lead to texture which are either "too smooth" or "too different" from the original texture. As for the perspective projection angle (such as defined in figure 4.8), it should be less than 70 degrees. In summary, the landmark can not be observed from a position which is either "too close", "too far away", or with a "large" projection angle to a landmark, according to the above experimented quantities.

As expected recognition performed best for the specialized landmark due to its scalability which allowed to very small templates to successfully match the observation, and to very big templates to still represent a faithful texture (due to its regularity and smoothness). The landmark representing a blackboard corner performed worse than the specialized landmark but much better than the one representing the computer corner. Figure 4.18 bottom-row right-hand shows the resulting shapes of an approximate observation areas for three landmarks including acquisition and observation points.

Summary

The results of the first experimentation phase were very encouraging because they demonstrated the advantage of proposed recognition method based on reprojection of reference-views, which can be quantified in terms of a "wide" observation-area. In particular, compared with other methods which directly re-use the acquired landmark appearance as visual prediction for the matching, the proposed method open up to a much larger recognition possibility. Landmarks can in fact still be recognized also when current observation and acquisition viewpoints are relatively distant, (i.e. current viewpoint does not allow for a direct re-use or acquired texture patches for the match). The advantage consequently is a more flexible recognition, in other words, a much wider "observation-area".

It is expected that performance would decrease when estimated robot and landmark pose will be obtained by an automatic process, and this is the reason why observation areas were thoroughly tested in a later experimentation (presented in chapter 6). It was important, nevertheless, to perform the first experimentation in "favorable" conditions in order to achieve a base for comparison and so understand how far can we get in term of performance with next proposed experimentation.

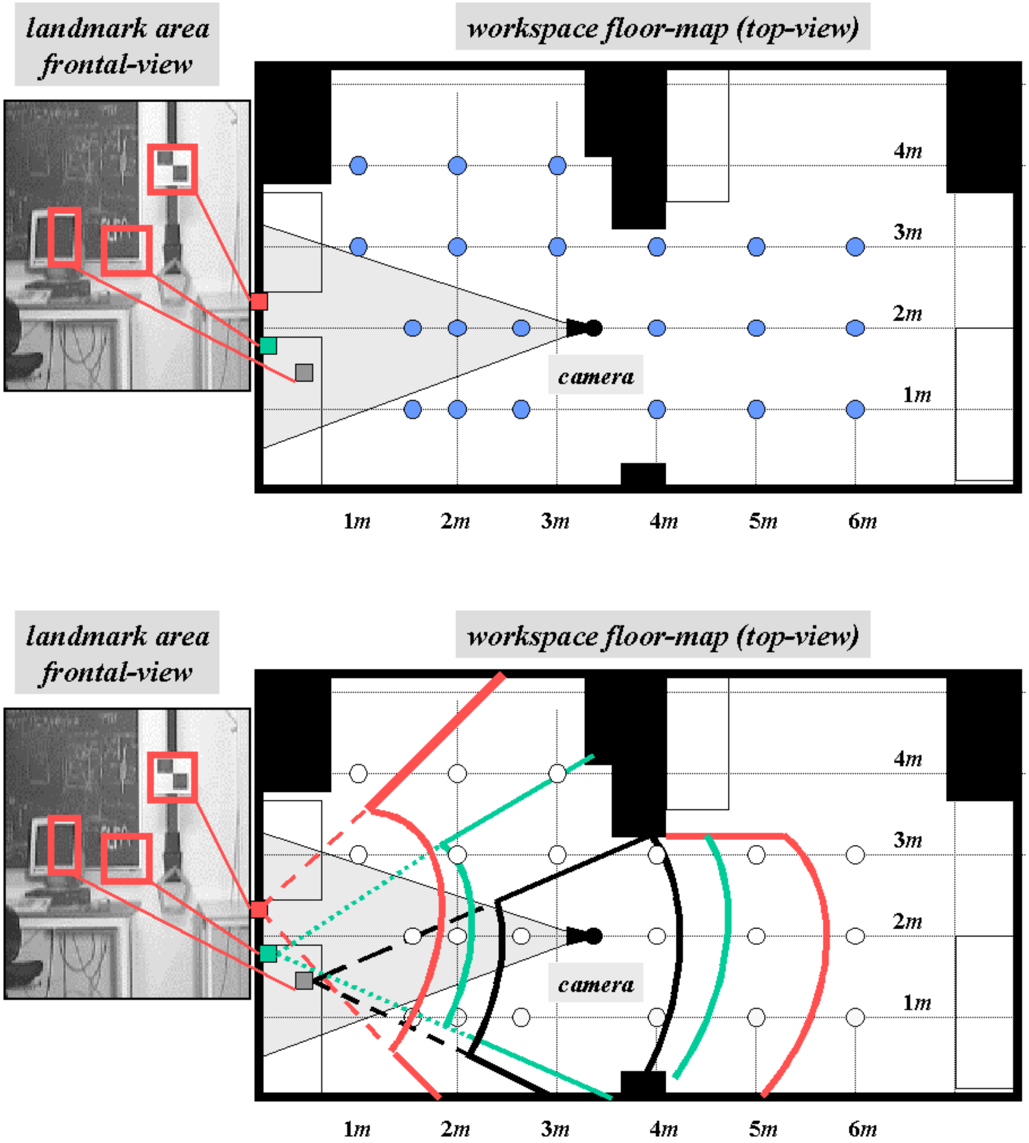

landmark area frontal-view
workspace floor-map (top-view)
4m
3m
2m
1m
camera
1m
2m
3m
4m
5m
6m
landmark area frontal-view
workspace floor-map (top-view)
4m
3m
2m
1m
camera
1m
2m
3m
4m
5m
6m

Figure 4.18: The figures left-hand show the reference views associated to the three landmarks. The figure top-right shows a floor-map containing the grid of acquisition and observation points. The figure bottom-right shows the resulting shapes of an approximate observation areas for three landmarks including acquisition and observation points.

4.2 Triangulation-based Robot Localization

The accuracy and reliability of a localization process based on external sensors can be improved by an effective exploitation and combination of sensed information. This "belief" is based on the fact that sensors like vision perform well on humans and animals, where the provided information is exploited and combined for the purpose of localization. This fact has pointed out that developing more efficient localization algorithms based on vision and combination of sensed information is a field with a dramatic potential.

In the literature the exploitation of multiple sensory readings is often been proposed in order to: (1) validate reliability of a single sensory input; (2) robustly estimate positional information.

In the first case, the analysis of the reliability of a single sensory input is based on an expected data coherence. For example, the comparison of consecutive readings related to the same environment feature, such as in tracking operations, (Leonard and Durrant-Whyte [82], Jensfelt and Christensen [74], Davison and Murray [41]), or the comparison of "simultaneous" readings related to different feature parts, such as for example in a sonar-array response to a wall, (Sabatini, [121]).

In the second case, the robust estimate of positional information is based on the combination of multiple measurements, (Carlsson [22], Durrant-Whyte and Leonard [48]). For example, a combination of "simultaneous" sensory readings related to one or several parts of the environment in a triangulation scheme, based on vision (Atiya and Hager [7], Feng et al. [52], Andersen and Gonçalves [3], Madsen and Andersen [97]), ultrasonic range-finders (Di Benedetto and Livatino, [10], Sabatini, [121]), etc.

In the second case a great advantage is often represented by the possibility of an *absolute* estimation of robot position and heading, (instead than a relative estimation). An absolute pose estimate can in fact allow the system to "erase" previously arisen errors, and thus prevent errors from accumulating. It is in the second case this thesis focuses in order to estimate an absolute, accurate and reliable robot-pose. The argumentation for the use of of triangulation have been presented in subsection 3.1.2.

4.2.1 Robot Centered Triangulation

The theoretical aspects of obtaining the robot position and heading from triangulation have been demonstrated from various researchers, (Andersen and Gonçalves, [3], Betke and Gurvits, [12], Atiya and Hager [7], Feng et al [52]).

The triangulation method such as presented in Madsen and Andersen, [97], estimates camera position, i.e. the position of the optical center of the camera, and camera orientation, i.e. the direction of the optical axis relative to the world coordinate system. By knowing camera position and orientation relative to robot center, the camera position can be converted to robot position. Analogously, using knowledge of the camera pan angle relative to the robot the direction of the camera optical-axis can be converted to robot heading.

The basic idea in the robot centered triangulation is to consider the robot as the local sensing system. The robot senses the landmarks by measuring the angles to them with respect to its reference axis. This process allows for measurements in a local moving coordinate system. This local frame is possibly rotated and translated with respect to the global fixed coordinate system, within which the position of landmarks are known. This is represented in figure 4.19. The robot position and orientation can then be estimated by determining the transformation between the local coordinate system and the global one.

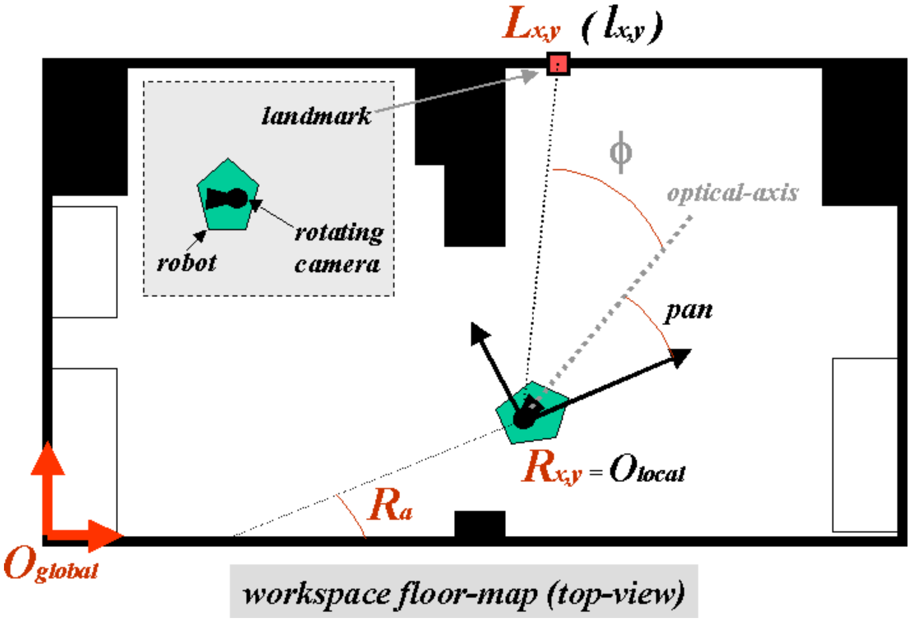


Figure 4.19: Robot centered position estimation. The robot is rotated R_α with respect to the global coordinate system and positioned at R_x, R_y, viewing a landmark at an angle $l_\tau = pan + \phi$.

The proposed system is intended to be used in indoor semi-structured environments. Consequently, the robot is only moving around on a planar surface, which makes sufficient that the robot localizes itself by estimating its 2D position and heading.

The robot is equipped with a single camera that can freely pan 360° relative to the robot heading. The pan angle relative to the robot heading can be obtained from an encoder in the pan motor. The local frame origin is the camera optical-center, the local frame orientation is the robot orientation. The camera orientation to the robot is the measurable pan angle.

We now present the formulae needed for triangulating the robot position and heading based on an image (obseved view) containing three landmarks at known locations. The position of the three landmarks in world coordinate system is also known. The main parameters are represented in figures 4.19 and 4.20 are described in the following:

- the *robot center* in global coordinates, denoted R_x and R_y, and heading denoted R_α;
- the *landmark center* in global coordinates, denoted L_x and L_y;
- the *landmark center* in local coordinates, denoted l_x and l_y;
- the *landmark center* in the camera image-plane, denoted $W = (W_u, W_v)$;
- the *angle to a landmark* as deviation from the optical-axis, denoted ϕ;
- the *pan angle*, i.e. the angle of the camera optical-axis in local coordinates, denoted pan;
- the *angle to a landmark* in local coordinates, denoted $l_\tau = \phi + pan$;
- the *camera focal length* in pixels, denoted f.

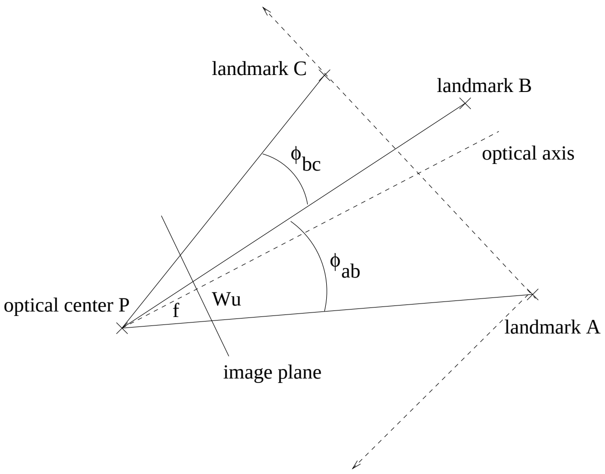


Figure 4.20: Top view of pin-hole camera and lines of sight for the three landmarks. The angular separations, ϕ^{ab} and ϕ^{bc}, form the basis for computing optical-center position, P.

Knowing f and W, it is possible to measure:

$$\phi = arctan\left(\frac{W_u}{f}\right) \tag{4.3}$$

so that

$$l_\tau = pan + \phi \tag{4.4}$$

Form the tangent relation it is known that

$$tan(l_\tau) = \left(\frac{l_y}{l_x}\right) \tag{4.5}$$

Knowing that the rotation and translation transformation of a point from a global 2D coordinate frame to a local frame is given by the equations 4.6.

$$\begin{aligned} l_x &= (L_x - R_x)\ \ cos(R_\alpha) + (L_y - R_y)\ \ sin(R_\alpha) \\ l_y &= (L_y - R_y)\ \ cos(R_\alpha) - (L_x - R_x)\ \ sin(R_\alpha) \end{aligned} \tag{4.6}$$

Combining equations 4.5 and 4.6:

$$tan(l_\tau) = \frac{(L_y - R_y) - (L_x - R_x)\ \ tan(R_\alpha)}{(L_x - R_x) + (L_y - R_y)\ \ tan(R_\alpha)} \tag{4.7}$$

In equation 4.7 the unknowns to be determined are R_x, R_y and R_α. The other parameters are instead known for each landmark. That is, the position of the landmark in the global reference frame (L_x,L_y) and the measured angle to a landmark l_τ.

Applying equation 4.7 for a set of landmarks will thus constrain the possible solutions. For three landmarks the equation is fully determined since we then have three equations with three unknown, (and nine known parameters). For more landmarks a least squared solution may be obtained (reducing the uncertainty in the pose estimate).

Let us now consider three landmarks denoted a, b and c, with corresponding image locations W^a, W^b, and W^c. Assuming the W_u-axis of the image plane is parallel to the plane of motion of the robot, the angle between the optical axis and the line of sight to the landmark can be computed as in the following:

$$\phi^a = arctan\left(\frac{W_u^a}{f}\right) \tag{4.8}$$

$$\phi^b = arctan\left(\frac{W_u^b}{f}\right) \tag{4.9}$$

$$\phi^c = arctan\left(\frac{W_u^c}{f}\right) \tag{4.10}$$

Consequently, the angular separations can be computed:

$$\begin{aligned} \phi^{ab} &= \phi^a - \phi^b \\ \phi^{bc} &= \phi^b - \phi^c \end{aligned} \tag{4.11}$$

Figure 4.20 shows these angles along with other variables.

Based on the angular separations the camera position is computed in a coordinate system defined relative to the three landmarks. In particular, the three landmarks define a local coordinate system with origin in the landmark which lies on the far right in the image plane, (landmark a in considered case), and x-axis through landmark which lies on the far left, (landmark c in considered case). This coordinate system is called *landmark coordinate system*. Figure 4.20 defines the landmark coordinate system and other variables.

In the following a notation is used, where for example $\angle_{abc}$ denotes the angle between the line from landmark b to a, and the line from landmark b to c. Analogously, the $\angle_{Pab}$ is the angle between the line from the optical center P to landmark a, and the line from landmark a to b. Additionally, $dist_{ab}$, denotes the distance between landmark a and b.

Based on Andersen and Gonçalves, [3], there are two landmark topologies, topology 1 and topology 2, defined by the middle landmark, (b in our case), being respectively in front of or behind the line joining the other two. In case of a *collinear* configuration of landmarks the two topologies become equivalent. Figure 4.20 shows a topology 2 landmark configuration. Note that the landmark-coordinate system is independent of topology.

Two angles needed to compute the camera position depend on the topology. These angles are denoted $\angle_{Pab}$ and $\angle_{bcP}$, and they are computed as presented in the following.

- **Topology 1**

  Using $\phi^{ac} = \phi^{ab} + \phi^{bc}$:

$$\angle_{Pab} = -\arctan\left(\frac{dist_{bc}\sin(\phi^{ab})\left(\sin(\phi^{ac})\cos(\angle_{abc}) - \cos(\phi^{ac})\sin(\angle_{abc})\right)}{dist_{ab}\sin(\phi^{bc}) + dist_{bc}\sin(\phi^{ab})\left(\cos(\phi^{ac})\cos(\angle_{abc}) + \sin(\phi^{ac})\sin(\angle_{abc})\right)}\right) \tag{4.12}$$

$$\angle_{bcP} = -\angle_{Pab} - \phi^{ac} + \angle_{abc} \tag{4.13}$$

- **Topology 2**

$$\angle_{Pab} = \arctan\left(\frac{dist_{bc}\sin(\phi^{ab})\left(-\sin(\phi^{ac})\cos(\angle_{abc}) - \cos(\phi^{ac})\sin(\angle_{abc})\right)}{dist_{ab}\sin(\phi^{bc}) + dist_{bc}\sin(\phi^{ab})\left(\cos(\phi^{ac})\cos(\angle_{abc}) - \sin(\phi^{ac})\sin(\angle_{abc})\right)}\right) \tag{4.14}$$

$$\angle_{bcP} = b\pi - \angle_{Pab} - \phi^{ac} - \angle_{abc} \tag{4.15}$$

When the angles $\angle_{Pab}$ and $\angle_{bcP}$ have been computed according to the correct topology, the position of the camera can be computed relative to the landmark-coordinate system, (with origin in the landmark a), shown in figure 4.20, by the following equations.

$$cam_x = dist_{ac}\,\frac{\tan(\angle_{bcP} + \angle_{acb})}{\tan(\angle_{Pab} + \angle_{bac}) + \tan(\angle_{bcP} + \angle_{acb})} \tag{4.16}$$

$$\tag{4.17}$$

$$cam_y = dist_{ac}\,\frac{\tan(\angle_{bcP} + \angle_{acb})\tan(\angle_{Pab} + \angle_{bac})}{\tan(\angle_{Pab} + \angle_{bac}) + \tan(\angle_{bcP} + \angle_{acb})} \tag{4.18}$$

The camera coordinates cam_x and cam_y relative to the landmark-coordinate system can be transformed to global world coordinates, (R_x, R_y), using the location of the origin of the landmark coordinate system, i.e., (L_x^a, L_y^a). Let μ be the deductible angle between the world-coordinate x-axis and the landmark-coordinate x-axis, then:

$$\begin{aligned} R_x &= cam_x \quad \cos(\mu) - cam_y \quad \sin(\mu) + L_x^a \\ R_y &= cam_x \quad \sin(\mu) + cam_y \quad \cos(\mu) + L_y^a \end{aligned} \tag{4.19}$$

and

$$R_\alpha = arctan \left(\frac{L_y^a - R_y}{L_x^a - R_x} \right) - (pan + \phi^a) \tag{4.20}$$

In summary:

- the **robot position**, (R_x, R_y), is function of the angular separations and the positions of the landmarks in the world model:

$$\begin{aligned} R_x &= \text{function}(\phi^{ab}, \phi^{bc}, L^a, L^b, L^c) \\ R_y &= \text{function}(\phi^{ab}, \phi^{bc}, L^a, L^b, L^c) \end{aligned} \tag{4.21}$$

- the **robot heading**, R_α, is a function of pan angle, the angular separations between landmarks, and landmark world positions.

$$R_\alpha = \text{function}(pan, \phi^{ab}, \phi^{bc}, L^a, L^b, L^c) \tag{4.22}$$

4.2.2 Problem and Sensitivity Analysis

Naturally, when using real data the estimation of robot position and heading will be error prone. As described in the previous section several parameters influence the result of a triangulation-based self-localization process. For example, the camera focal-length, f, must be very accurately calibrated if we want to make sure its uncertainty would not cause an error in the recovered positions. Analogously, the landmark pose, $L_{(x,y,z,\theta)}$, should be absolutely correct if we want to avoid that this inaccuracy causes errors, etc. Proposing a method for robot pose estimation is consequently a challenging topic since much attention have to be paid on the resulting accuracy, reliability, and error propagation.

The error in robot pose may arise from inaccuracies in knowledge provided prior to robot navigation, e.g. environment model, camera focal length, etc., and in the knowledge provided during robot navigation, e.g. sensor measurements, current pose estimates, etc. In addition, the localization problem is also challenged by the dynamic nature of typical indoor environments, and by the fact that the environments have not been specially structured for the purpose of robot localization. In general, the many sources of error may be distinguished as:

- *Systematic errors.* For example those caused by inaccuracies in: landmark position and orientation, robot initial pose (if provided), camera focal length.

- *Non-systematic errors.* For example those arising during navigation caused by inaccuracies in: sensor measurements, (e.g. odometric errors), robot current and previous pose-estimates, algorithmic errors (numerical imprecision), etc.

The many significant sources of inaccuracy represent a major challenge for reasoning with uncertainty in mobile robotics because the involved aspects are difficult to model as stochastic perturbation process. In this thesis it is therefore assumed that focal length and pan-angle are correct and numerical imprecision neglectible. Rather, the focus will be on how errors in angular separation between a landmark-triplet, i.e. ϕ^{ab} and ϕ^{bc}, and landmark global positions, i.e. errors in $L^a_{(x,y)}$, $L^b_{(x,y)}$, and $L^c_{(x,y)}$, propagate to uncertainty in robot pose estimate, R_x, R_y, and R_α.

As seen in subsection 4.2.1 the robot position and heading are functions of pan-angle, angular separations, and landmark positions. Taking into account the parameters that we assume to be correct, the error affecting robot pose may be due to uncertainties in estimated angular separations between landmarks, ϕ^{ab}, ϕ^{bc}, and landmark world positions $L^a_{(x,y)}$, $L^b_{(x,y)}$, and $L^c_{(x,y)}$. Being interested in 2D localization (the robot navigates on a planar surface), the uncertainty in L_z can be neglected. As for the uncertainty in L_θ, this can be neglected if landmarks have been selected with a "convenient" orientation to camera.

The effect of uncertainties on angular separations (arising from uncertainty in landmark locations W_u), and landmark position, is sketched in figure 4.21. As it is possible to observe, when the pose is estimated based on the triangulation method, the effect of uncertainty in considered input parameters is expected to propagate to uncertainty in robot pose, $R_{(x,y,\alpha)}$.

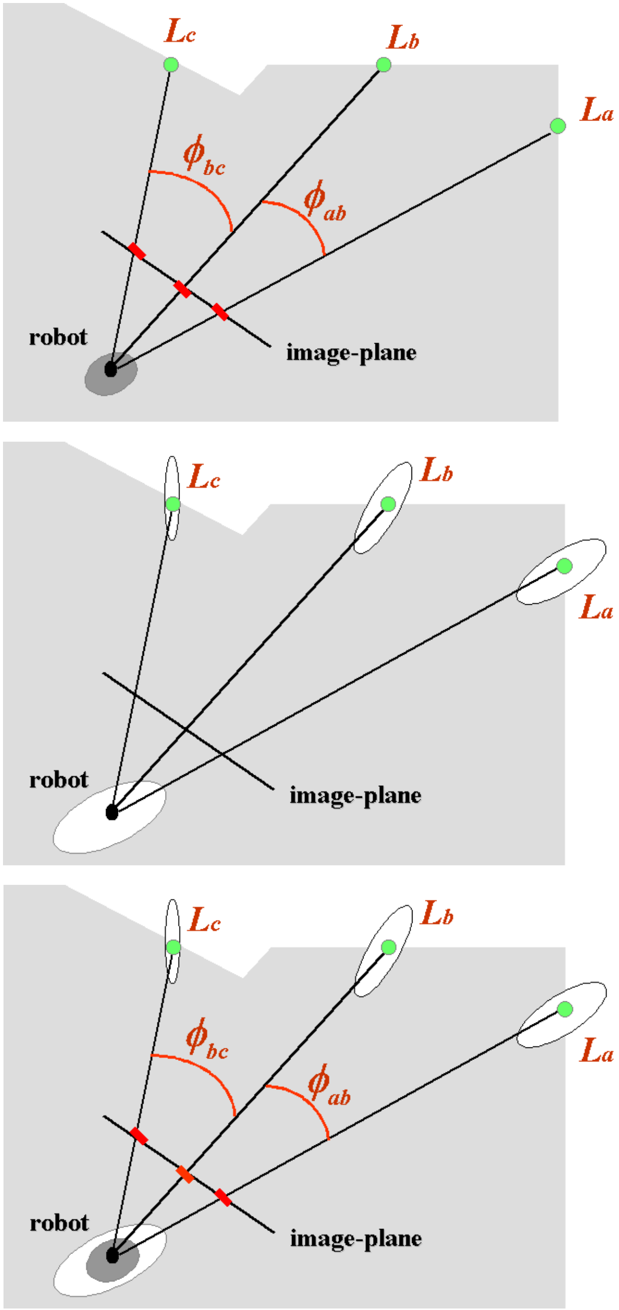


Figure 4.21: The figures schows through synthetic examples the effect of: (1) an angular separation error (short red segments) arising at navigation time, due to template matching, (top-row); (2) a landmark pose error arising at learning time (white ellipses), due to 3D reconstruction uncertainty (middle row). These errors will be someway combined (bottom-row), and depending on the current robot position and other factors one of them could prevail on the other.

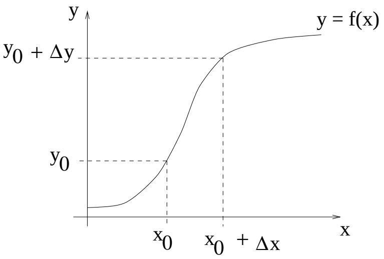


Figure 4.22: A function of one variable and the increment function value resulting from an increment in variable.

Sensitivity Analysis

For the purpose of analysis the parameters W_u^i and $L_{(x,y)}^i$, $i = a, b, c.$, shall be considered subjected to additive Gaussian noise of zero mean and variance respectively $\sigma_{W_u}^2$, $\sigma_{L_x}^2$, $\sigma_{L_y}^2$. The Gaussian model is just a vehicle for a compact analysis. We are interested in the sensitivity towards errors, not in the particular distribution model. The sensitivity analysis is based on first order covariance propagation, [96, 97], that is, propagating the variances $\sigma_{W_u}^2$, $\sigma_{L_x}^2$, $\sigma_{L_y}^2$, to the 3×3 covariance matrix for robot 2D position and 1D heading, (R_x, R_y, R_α).

The proposed technique used in the sensitivity analysis is known as *Covariance Propagation*, (Haralick [65]), demonstrated successfully in numerous applications. The technique is basically a generalized analysis of the error propagation and offers a compact and complete way of formulation to how the noise in a set of input parameters affects a set of output parameters for some computation. In our case the computation is the robot position and heading from a robot-centered triangulation.

In order to introduce the principle behind the method, let us consider a function of one variable, $y = f(x)$, such as for example, the function represented in 4.22. A change Δx, in the input variable causes a change in the output variable. Using Taylor expansion:

$$y_0 + \Delta y \quad = \quad f(x_0) + \Delta x \cdot f'(x_0) + (\Delta x)^2 \cdot f''(x_0)/2! + \cdots \tag{4.23}$$

Since $y_0 = f(x_0)$, Δy can be defined as:

$$\Delta y \quad = \quad \Delta x \cdot f'(x_0) + (\Delta x)^2 \cdot f''(x_0)/2! + \cdots$$

Thus, in a first order approximation $\Delta y = \Delta x \cdot f'(x_0)$. If the input variable is subject to additive noise of zero mean, Δx could be considered a random variable of some variance and zero mean, and thus the first order approximation gives the resulting noise in the output parameter Δy.

It should be noted that the first order approximation to the Taylor expansion is computed around an "operating point", x_0. Thus, the effect of the input noise varies depending on which neighborhood of the input parameter is being studied.

C.B. Madsen, [96], demonstrated viewpoint variation of noise sensitivity in cases where viewpoint defined the operating point. The operating point is in our case represented by camera poses and landmark locations in the image-planes. Consequently, other than estimating the uncertainty the goal of the sensitivity analysis also is to understand the relationships between landmark reconstruction technique and current robot-landmark configuration.

In terms of the one-dimensional example if the noise in the input parameter has standard deviation σ_x, then the resulting standard deviation in the output parameter noise is $\sigma_y = \sigma_x \cdot f'(x_0)$. The covariance propagation method is a generalization of this procedure into multiple dimensions involving both variances and covariances of parameters.

The technique described above represents for the proposed method a suitable way of estimating propagation of error in the input parameters, and so this is proposed is order to characterize the uncertainty in computed landmark positions.

Because the robot is moving on a planar surface, the z component of robot position is not considered when computing the error propagation, (R_z is assumed constant and a priori known). Consequently, the input variable of interest when computing robot position and heading by the proposed triangulation method, can be described by $\vec{q} = [\phi^{ab}, \phi^{bc}, L_x^a, L_y^a, L_x^b, L_y^b, L_x^c, L_y^c]$. If we call h the function used to compute robot position and heading, the resulting robot position and heading $\vec{R} = [R_x, R_y, R_\alpha]$, can be obtained by the equation:

$$\vec{R} \ = h \ (\ \vec{q} \ ) \tag{4.24}$$

The relationships between $\vec{R}$ and $\vec{q}$ are described by the following equations (equivalent to 4.19 and 4.20).

$$\begin{aligned} R_x &= cam_x \quad \cos(\mu) - cam_y \quad \sin(\mu) + L_x^a \\ R_y &= cam_x \quad \sin(\mu) + cam_y \quad \cos(\mu) + L_y^a \end{aligned} \tag{4.25}$$

and

$$R_\alpha = arctan\left(\frac{L_y^a - R_y}{L_x^a - R_x}\right) - (pan + \phi^a) \tag{4.26}$$

where: $(cam_x, cam_y$ represents the camera position relative to the landmark-coordinate system, (L_x^a, L_y^a) the origin of the landmark coordinate system, μ the angle between the world-coordinate x-axis and the landmark-coordinate x-axis, and (R_x, R_y, R_α) the robot pose in global-coordinate system.

Under a first order approximation the covariance matrix Υ for the robot pose $\vec{R}$ related to $\vec{q}$ becomes:

$$\Upsilon \; = \; \frac{\partial \vec{R}}{\partial \vec{q}} \; \Lambda \; \frac{\partial \vec{R}}{\partial \vec{q}}^{T} \tag{4.27}$$

where Λ is the covariance of the input variables of interest, (i.e. the variables contained in $\vec{q}$). The equation 4.27 represents how the uncertainty in $\vec{q}$ (as expressed by Υ) can be propagated to the robot pose parameters $\vec{R}$.

Because of the many uncertain parameters involved in the sensitivity analysis, it is proposed to first only consider ϕ^{ab}, and ϕ^{bc}, as the uncertain parameters, (so assuming $L^{a}_{(x,y)}$, $L^{b}_{(x,y)}$, $L^{c}_{(x,y)}$, being correct), and analyze the effect of these uncertainties in the "triangulated" robot pose. Then, it is proposed to only consider $L^{a}_{(x,y)}$, $L^{b}_{(x,y)}$, $L^{c}_{(x,y)}$ as the uncertain parameters, (so assuming W^{a}_{u}, W^{b}_{u}, W^{c}_{u}, being correct), and analyze the effect of these uncertainties in the "triangulated" robot pose. Eventually, both the group of parameters will be considered as uncertain.

Angular Separation Between Landmarks

The angular separation is estimated based on the focal length, f, and the detected landmark locations, W_u, (see equation 4.3). Assuming f to be correct, an error on W_u may due to the algorithm used for estimating W_u, i.e. the template matching.

The detection of landmark image location based on template matching (normalized cross correlation) can provide a very accurate estimate but inaccuracy may arise as well. Madsen and Andersen [97] demonstrated how accurately template matching (normalized cross correlation) can locate landmarks in the image plane. In particular, it is shown that under good conditions image locations can be determined with a standard deviation ranging from approximately 1.2 pixels to 3.6 pixels. Experiments on real images showed that this noise is equivalent to a standard deviation of 2 pixels, in case of camera resolution of 512 × 512 pixels and a field of view of about 53°.

Inaccuracies in estimates of landmark image location, W_u, may propagate to uncertainties in the estimate of the angular separation between landmarks which may then propagate to uncertainties in robot position and heading.

The effect of this error in robot position and heading, estimated by the proposed self-localization method can be described as in equation 4.27, having the input parameters $\vec{q}$ = [W^{a}_{u}, W^{b}_{u}, W^{c}_{u}]. These three parameters are considered independent, so that Λ can be described as in the following:

$$\Lambda = \begin{bmatrix} \sigma^2_{W^a_u} & 0 & 0 \\ 0 & \sigma^2_{W^b_u} & 0 \\ 0 & 0 & \sigma^2_{W^c_u} \end{bmatrix} \tag{4.28}$$

and the matrix of the partial derivatives $\frac{\partial \vec{R}}{\partial \vec{q}}$, can be described by:

$$\frac{\partial \vec{R}}{\partial \vec{q}} = \begin{bmatrix} \frac{\partial R_x}{\partial W_u^a} & \frac{\partial R_x}{\partial W_u^b} & \frac{\partial R_x}{\partial W_u^c} \\ \frac{\partial R_y}{\partial W_u^a} & \frac{\partial R_y}{\partial W_u^b} & \frac{\partial R_y}{\partial W_u^c} \\ \frac{\partial R_\alpha}{\partial W_u^a} & \frac{\partial R_\alpha}{\partial W_u^b} & \frac{\partial R_\alpha}{\partial W_u^c} \end{bmatrix} \tag{4.29}$$

The resulting covariance matrix will then be as in the following:

$$\Upsilon = \begin{bmatrix} \sigma_{R_x}\sigma_{R_x} & \sigma_{R_x}\sigma_{R_y} & \sigma_{R_x}\sigma_{R_\alpha} \\ \sigma_{R_y}\sigma_{R_x} & \sigma_{R_y}\sigma_{R_y} & \sigma_{R_y}\sigma_{R_\alpha} \\ \sigma_{R_\alpha}\sigma_{R_x} & \sigma_{R_\alpha}\sigma_{R_y} & \sigma_{R_\alpha}\sigma_{R_\alpha} \end{bmatrix} \tag{4.30}$$

The effect in the estimated robot position of an error in locating landmarks in the image-plane is visually illustrated in the simulated example of figure 4.23. This result has been experimented by the author of this thesis in the real mobile robotic system described in section 4.1.5, (Livatino and Madsen [93]). This figure shows a snapshot from a particular position while the robot is navigating the old-laboratory along some path. To simulate the positional uncertainty accruing from inaccuracies in determining landmark image locations, noise has been added to landmark images locations.

The ellipse in figure 4.23, around the computed robot positions, represents the positional uncertainty for a given robot position and landmark triplet, after 1000 simulated runs. The ellipse in figure represents a very high uncertainty since the floor map depicts a 8m×5m room, (the old laboratory), so that the computed positions cover around 1.5m × 0.5m area. Such an uncertainty is prohibitive for navigation in cluttered environments and narrow passages like doorways and it can lead to errors which can be so severe that navigation is impossible. Figure 4.24 shows how the noise sensitivity varies drastically with the choice of landmarks. In particular, the effect of this noise is an error on the computed robot position, which may vary from few centimeters to many meters.

We can conclude that the computation of robot position and heading based on triangulation of landmarks can provide a very accurate estimate but it is shown to be very sensitive to noise in estimated angular separation between landmarks. The reason for the noise to arise is mainly due to spatial landmark configuration and relative position between robot and landmarks.

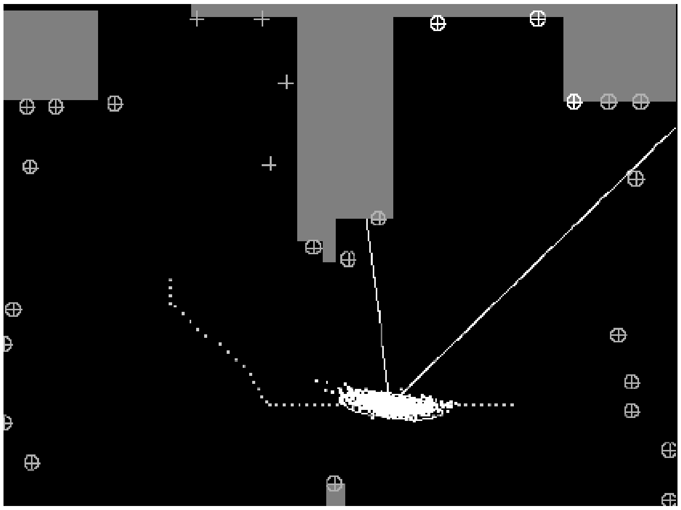


Figure 4.23: 2D floor map of robot workspace. Obstacles are shown in grey, free-space in black. The robot is following an assigned path, represented by small white points. All landmarks are shown as crosses. Visible landmarks are within a circle and those used for computing position have a white circle. The cloud of white points are robot positions computed with noise added to the sensory measurements (see text). The two white lines emanating from the robot position indicate the camera field-of-view. The ellipse is the positional uncertainty computed by the 1000 simulated runs.

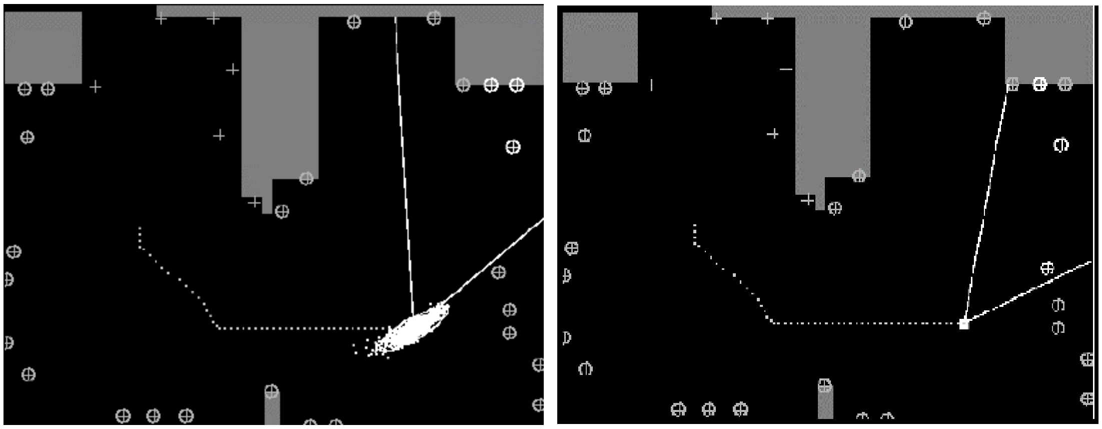


Figure 4.24: Variation in noise sensitivity when choosing different landmark triplets, but keeping the same robot position. In the right hand figure the chosen triplet provides a much better position estimate than that in the left figure.

Landmark World Position

The estimation of landmark world position may be performed manually or automatically during a learning phase (subsection 1.5.4). In case of manual measurements, the estimated landmark position is expected accurate and with the same level of uncertainty for all landmarks. In case of automatic computation, the estimated landmark position is expected less accurate and with a varying level of uncertainty depending on the acquisition system configuration. In particular, as it is demonstrated in chapter 5, the estimated landmark pose can provide a very accurate estimate but inaccuracy may arise as well. This uncertainty is mainly due to inaccuracies in camera pose, baseline length, distance between camera and landmarks, landmark vergence and orientation angle, (see sub-section 5.3).

Performed experimentations showed that under good conditions landmark position was determined with a standard deviation ranging from approximately 0 cm. to 3 cm. in the direction perpendicular to camera baseline, and 0 cm. to 1 cm. in the direction parallel to camera baseline. Consequently, depending on the relative spatial configuration between landmark and camera at acquisition time, the effect of an error in landmark world position X and Y may cause a different discrepancy from correct value in robot pose estimate.

Analogously, for what was proposed in case of an uncertainty in angular separation between landmarks, the effect of an error in landmark position to robot position and heading, estimated by the proposed self-localization method, can be described as in equation 4.27, having $\vec{q} = [L_x^a, L_y^a, L_x^b, L_y^b, L_x^c, L_y^c],$. These six parameters are considered independent for different landmarks, but not for the same landmark, so that Λ and Υ can be computed as in the following:

$$\Lambda = \begin{bmatrix} \sigma_{L_x^a}\sigma_{L_x^a} & \sigma_{L_x^a}\sigma_{L_y^a} & 0 & 0 & 0 & 0 \\ \sigma_{L_y^a}\sigma_{L_x^a} & \sigma_{L_y^a}\sigma_{L_y^a} & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{L_x^b}\sigma_{L_x^b} & \sigma_{L_x^b}\sigma_{L_y^b} & 0 & 0 \\ 0 & 0 & \sigma_{L_y^b}\sigma_{L_x^b} & \sigma_{L_y^b}\sigma_{L_y^b} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{L_x^c}\sigma_{L_x^c} & \sigma_{L_x^c}\sigma_{L_y^c} \\ 0 & 0 & 0 & 0 & \sigma_{L_y^c}\sigma_{L_x^c} & \sigma_{L_y^c}\sigma_{L_y^c} \end{bmatrix} \tag{4.31}$$

and according to Λ, $\frac{\partial \vec{R}}{\partial \vec{q}}$ and Υ can be estimated.

The problem was studied in a concrete context using real data. Experiments were performed on some sample triplets by adding noise to landmark pose according to the errors measured in real setting, (see subsection 5.2.4). In particular, three different types of landmark triplet were considered: (1) "close to ideal", i.e. topology 1 and landmark locations "spread" in the image-plane; (2) "spread" in the image-plane but almost collinear landmarks; (3) collinear landmarks and not "spread".

The effect of an error in landmark pose varied from "accurate" (around 2 cm) to "inaccurate" (above 10 cm), depending on the chosen landmark triplet. It was possible to conclude then that the effect of an error in resulting robot position (calculated by the triangulation method) mainly depends on the relative position between robot and landmarks, and landmark configuration. In particular, the error in robot position estimates varied from a few millimeters to around one meter. This result demonstrated that to achieve accurate landmark estimates is possible in case the system had a strategy for selecting "good" landmark triplets.

To thoroughly investigate the propagation of an error in landmark world position to robot pose, it is not among the goals of this thesis work. The aim of the investigation in this subject was only in order to understand which are most significant sources of errors affecting robot pose estimate and which are expected consequences in robot localization performance, when the proposed self-localization method is applied.

Both Angular Separation Between Landmarks and Landmark World Position

Analogously, the same uncertainty estimation scheme could be used in case when errors in both angular separation and landmark world positions are considered. In this case $\vec{q} = [\phi^{ab}, \phi^{bc}, L_x^a, L_y^a, L_x^b, L_y^b, L_x^c, L_y^c]$ and according to this will be Λ and Υ.

As above mentioned, such a way of estimating uncertainty does not always represent the real error function, (the error is not always a linear function, [97]). Nevertheless, this way can provide the system with an expectation of uncertainty related to a specific choice for landmark triplet which could be taken into account when establishing which landmark triplet should be considered for robot pose estimation.

The above presented sensitivity analysis may become very complicated, especially in this particular case. An alternative is then represented by the technique of estimating expected positional error in robot pose by simulating it, (adding noise to input parameters according to real measurements), see figures 4.23 and 4.24. This technique may reveal to be a very practical method to estimate error propagation. To investigate performance sensitivity to errors in both angular separation between landmarks and landmark world position is a topic for future research.

4.2.3 Optimal Landmark-Triplet Selection

Once the effect of the main uncertain factors affecting the robot pose estimate have been understood, the issue becomes how to best exploit landmarks associated information, (e.g. position, orientation, positional uncertainty, observation area, etc.), in order to generate accurate and reliable robot-pose estimates.

Figure 4.24 showed how the noise sensitivity varies drastically with the choice of landmarks. In particular, the effect of the noise in landmark angular separation and landmark position is an error in computed robot pose which may vary from a few millimeters to some meters. This clearly demonstrates the need for a strategy for selecting ”good” landmark triplets, which in turn requires a technique for evaluating the possible landmark triplets before using them for triangulation.

To select *good* landmark triplets represents a fundamental step towards increasing accuracy and robustness of the proposed triangulation method. It is consequently very important to understand which parameters among the many available should be considered, and how these parameters should be weighted for the triplet selection. In other words, to establish on which base a triplet is elected as *optimal*.

Assuming that triangulation based self-positioning is fairly accurate, and that the robot does not move large distances between each time it checks its position, then the last position found by triangulation plus information from odometry (dead reckoning) gives a good estimate of current robot position. From this estimate the navigation system can compute:

1. *Visible landmarks.* That is, which landmarks in the environment model should be visible from the current position. Figures 4.23 and 4.24 show visible landmarks within a circle (all landmarks shown as a cross).

2. *Robot Position in Observation Areas.* That is, which landmark observation areas contain the current robot position. Main parameters which play a role in this, are the perspective projection angle (*pp-angle*) and the distance to landmark. Figure 4.25 shows an example of a trajectory along which current robot position can be contained into different observation areas. In particular, robot position R_1 is contained in observation areas relative to landmark L_c and L_a; robot position R_2 is contained in all landmark observation areas; and robot position R_3 is contained in observation areas relative to L_b and L_a.

3. *Determine all Landmark Triplets.* That is, to determine all permutations of landmark triplets that are within field of view for various pan angles. This is possible using the known camera focal-length and field of view, and all the landmark positions.

4. *Determine Optimal Triplet.* That is, to determine among all permutations the landmark triplet which minimizes the robot positional error. A measure of expected accuracy in ”triangulated” robot pose can be estimated based on the uncertainty analysis from previous sub-section (4.2.2). The latter is repeated for each valid combination of landmark triplet, so that, the system can eventually choose the optimal triplet before using it for triangulation-based robot pose estimate.

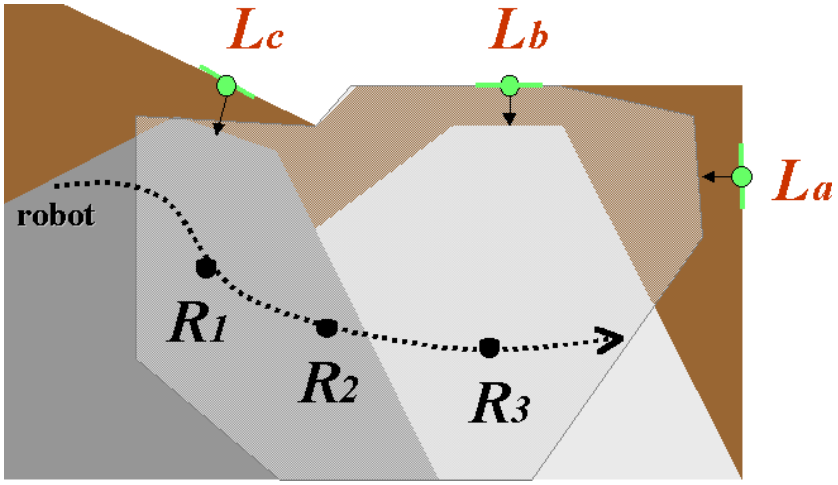


Figure 4.25: The figure shows an example of a trajectory along which current robot position can be contained into different observation areas. In particular, robot position R_1 is contained in observation areas relative to landmark L_c and L_a; robot position R_2 is contained in all landmark observation areas; and robot position R_3 is contained in observation areas relative to L_b and L_a.

Many different "goodness" measures can be relevant for various purposes. A thorough investigation of this is a direction for future research. In performed experiments, (presented in section 4.4), the applied measure is inversely proportional to the *area* of the uncertainty ellipse described previously, (Madsen and Andersen [97]). This measure simply results in a minimization of the region within which the computed positions will fall, when the landmark image locations are perturbed with noise.

An algorithm for selecting optimal landmark triplets which minimizes the error sensitivity, is regardless of whether the Gaussian model is applicable and regardless the fact that we are considering errors in angular separation or landmark pose. What changes in the case of angular separation or landmark pose is the parameters involved.

Analogously to what is proposed in case of errors in angular separation between landmarks, it is possible to predict the effect of errors in world landmark positions on the robot pose estimate, so that the system could choose the optimal triplet of landmarks also including this parameter.

Next experimentation section (4.4) will show results from navigated paths when an algorithm for optimal triplet selection is applied before robot pose is estimated with the proposed self-localization method.

4.3 Proposed Self-Localization Scheme

In the previous sections the theoretical aspects and argumentation for proposed landmark recognition and triangulation methods have been presented. This section is to describe how the proposed methods are applied to our system for the purpose of self localization. In particular, this section summarizes the required information and how the system uses and deals with them.

The a priori system knowledge is:

- camera focal length in pixels, (calibrated off-line);
- environment model, consisting of landmark reference images and positional information.

Unfortunately, the knowledge provided before the navigation does not suffice as the only information the system should rely on for self-positioning during navigation. The robot in fact needs to monitor its pose during the navigation as well as observe the current status of the environment. It is then proposed the use of the following external and internal sensory input:

- image observation, captured by the on-board vision system;
- odometric readings, continuously provided by robotic platform wheel encoders;
- pan-unit readings, provided by encoders in the pan motor.

Having the above described sensory input as well as the described priori knowledge, the procedure for visually determining the robot position and heading is as follows:

1. **Landmark Optimal-Triplet Selection**

   In order for the triangulation method to work, three landmarks must be observed by the robot on-board camera. This means that three landmarks should be selected such to allow the system to observe them in the same image from current robot position. The reason for having three landmarks in the same image-plane have been discussed in subsection 3.1.2.

   Having an estimate of current robot position, knowing camera focal length in pixels and landmark position and orientation in the workspace, the system is able to select the optimal triplet of landmarks, i.e. the triplet which minimize expected robot positional uncertainty (subsection 4.2.3). If the uncertainty associated to robot and landmark positional information as well as an estimate of landmark observation area, are available, the selection of optimal triplet can be more accurate.

   Knowing the camera pan-angle relative to the robot, the system is also able to compute the angle the pan-unit needs to turn in order to allow the system to observe the triplet by the ”most suitable” observation condition. The ”most suitable” observation condition for a landmark triplet is defined as the one making the two extreme landmarks in the camera image-plane, (according to the horizontal axis), at the same distance from the optical center.

2. **Landmark Detection**

   In order for the triangulation method to work, the location of the three selected landmarks in the current camera-view must be known. The selected landmark triplet represents the input to the landmark detection process which is then used for computing the landmark locations in the current image-plane.

   Having an estimate of the robot position and heading, knowing camera focal length in pixels, together with landmark position, orientation and reference images, the system is capable of locating the three selected landmarks in the image-plane based on the method demonstrated in section 4.1.

   In particular, landmarks can be detected during navigation in incoming images by first perspectively transforming, or re-mapping, reference landmark views to how they should appear in the current camera-view, and by then using the re-mapped images as template in a normalized cross-correlation procedure, in order to detect the landmarks in the incoming images.

3. **Robot Pose Estimation**

   The detected locations of landmarks in the image-plane represent the input to the triangulation method, (section 4.2), which is then used for computing the absolute camera position and heading (relative to the global coordinate system). This is possible since camera focal length in pixels, landmark positions as well as image locations, are known. Once the camera position and orientation has been estimated, this can easily be converted to robot position and heading, (knowing camera position relative to the robot).

   If the entire self-localization process is taking place while the robot is moving, the time elapsed between each robot pose computation should also be considered, hence, the robot pose should be updated by also taking into account the distance traveled during the time it took to run the localization process. This distance information can be achieved by a dead-reckoning system (odometry).

   The accuracy provided by the proposed localization algorithm based on triangulation is expected to be much higher than the odometry. However, the triangulation algorithm does not use information about the motion of the mobile robot, the history of position estimates, the commands that make the robot move, and the uncertainty in these commands. The system consequently should use a Kalman filter to update pose estimates and sets the odometry to the new filtered estimate, every time a landmark triplet is recognized.

   The proposed approach to global localization prevents errors from accumulating, hence, enables the system to run autonomously for a long time.

In summary, *landmark optimal-triplet selection* is to select 3 landmarks from the environment model such that they minimize robot positional error (and they are visible in one image); *landmark detection* is to detect locations of selected landmarks in the image-plane of current view; and *robot pose estimation* is to estimate robot position and heading by proposed triangulation technique based on detected landmark locations, (and odometric readings). The main computational steps together with main parameters required by the self-localization process, are depicted in figure 4.26.

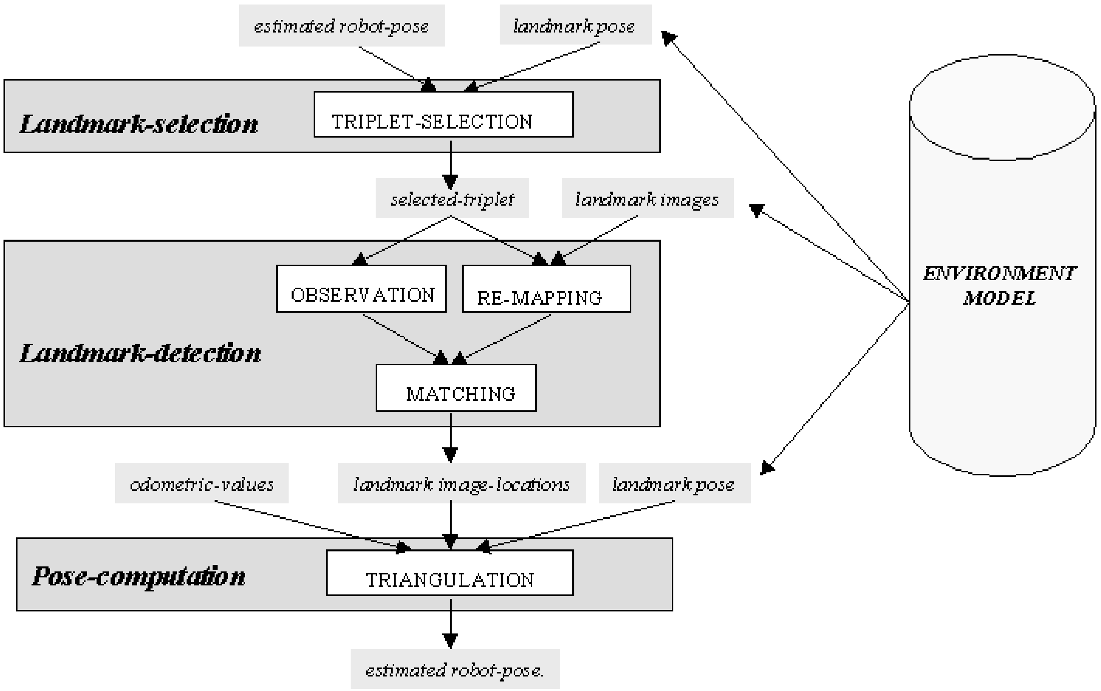


Figure 4.26: A schematic representation of the proposed self-localization procedure. The figure represents main computational steps and required parameters.

Automatic localization is only possible if the error in robot position during navigation is small enough to allow for recognition of the predicted landmarks triplet in the image plane. This possibility is unfortunately challenged by the many sources of uncertainty due to the navigation and feature measurements. Further to this, the complexity of indoor environments, and their dynamic nature, the short distances which might occur between obstacles, etc., call for a system capable of continuously monitoring self-localization. This would in fact minimize the intervals when the robot would either run ”blind” or with a high uncertainty, and would thus allow for reaching the goals of accurate navigation. It is consequently important that the sequence of steps described in figure 4.26 and commented above, run as fast as possible. This in fact will provide several advantages. Among them: the odometry will run for short time which allows for more accurate reading, and the size of search window can be small. Note that a faster landmark recognition may also involve a less precise landmark detection, this aspect consequently also needs attention.

4.4 Experimentation

A series of experiments was carried out in order to evaluate performance of the localization method described in the previous sections in a realistic setting. The performance of proposed localization method needed to be tested in a typical indoor environment in order to understand about potential, limits and possible improvements. In particular:

- how accurate is the estimated robot pose?
- how reliable is a navigation based on proposed localization method?
- how improved is the robot pose estimate using proposed triangulation method when compared to alternative solutions such as "odometry only" and "angular deviation to one landmark" [36].

It was also relevant to examine the limits of the main parameters involved in robot pose computation. For example, how accurate the "previous estimation" of robot pose needs to be.

The proposed method was implemented on a real robotic system (presented in subsection 4.1.5 and showed in figure 4.10) and experimented in the workspace previously described as the "old laboratory", (a room 8m x 5m floor size with plenty of furniture, shown in figure 4.27). Images of landmark reference views were provided and stored in a database, together with associated landmark positional information. Landmarks were related to objects both naturally occurring or on purpose introduced in the environment for the purpose of performance evaluation, and they were manually learned by the system, as described in 4.1.2. Figure 4.27 shows a typical trajectory followed to learn some of the landmarks.

The experimentation concerning self-localization was based on letting the robot navigate along a path of approximately 7 meters, which also included two rotations of approx. 45 degrees each. Figure 4.31 shows a typical path on the laboratory floor-map. The given path represented a trajectory that robot had to run along in order to get the laboratory exit-door. By considering the assigned path, 10 cm final error was considered acceptable. In fact, this would allow the robot to run through the door, thus giving a clear and intuitive understanding of the final error.

As previously described, the proposed localization scheme, based on 2D salient regions for landmarks, goes through three main computational steps: optimal triplet selection, landmark detection and robot pose computation, (section 4.3).

The robot was not supposed to stop during a run, so the system had to execute the proposed self-localization process while the robot was moving. The robot-pose estimate resulting from the triangulation was then updated by also including the distance traveled by the robot while the localization process was running (see section 4.3). This distance was obtained from odometric measurements. In particular, the distance traveled from previous pose estimate was considered.

The problem of positioning and updating speed are coupled. If positioning is more accurate then updating can be made less frequent. However, it is difficult to find a trade off between estimated accuracy and updating speed, since the error in the odometric system is unpredictable and depends on the followed trajectory. In particular, it is difficult to compute how

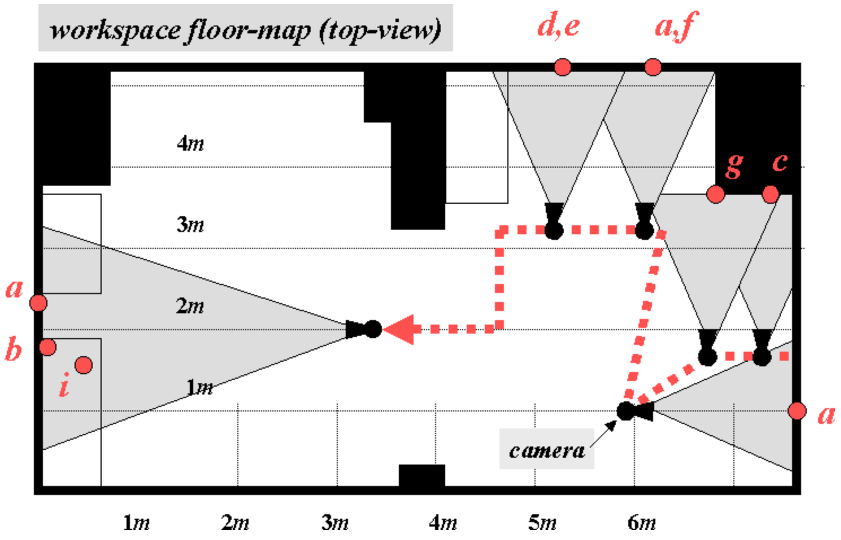


Figure 4.27: The figure shows a floor-map of the workspaces used in the experiments, (the old laboratory). The figure also shows the acquisition trajectories followed during the manual training together with camera poses at the moment landmarks were acquired.

far the robot can run before the error is too large to allow observation, in the same image, of the three predicted landmarks.

In order to overcome this problem, it was thought about using an high updating speed so that the error is kept small along any trajectory. A realistic estimate was to update the position up to 40-50 times in one run. However, in order to achieve this rate when using natural landmarks, it needed to optimize the algorithm for cross-correlation and to use a faster hardware. Consequently, specialized landmarks were used for some of the runs aimed to reach the above mentioned 40-50 times updating rate in one run.

Figure 4.28 left-hand shows a specialized landmark. This landmark allowed us to reduce the template image to a very small image, (the landmark center), as shown in figure 4.28 right-hand. Reducing the template size speeded up the localization process and allowed us to reach the desired updating rate with the available hardware and algorithms. The proposed localization method was also experimented successfully for shorter paths, (and lower update-rates), with visual landmarks which are naturally occurring, e.g., door handles, posters, light switches, etc. (see figure 4.29).

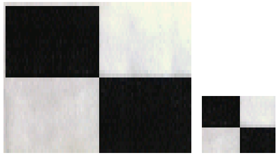


Figure 4.28: An image of the specialized landmark and its template image.

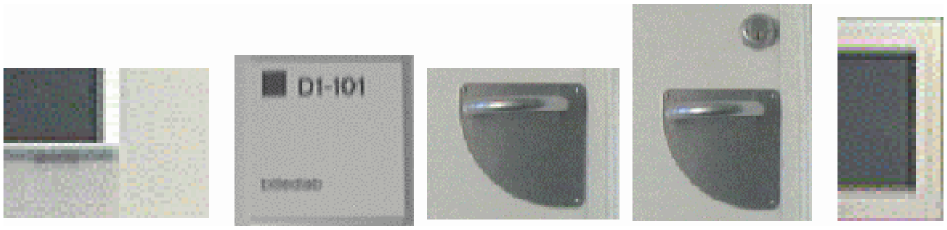


Figure 4.29: The figure shows visual landmarks which are naturally occurring, (door handles, posters, light switches, etc.)

During the experimentation it was important to perform many runs which required an accurate initial estimate. So, it was necessary to design a procedure in order to initialize the robot position automatically. The proposed procedure for initializing the robot position was based on recognition and triangulation of three colored landmarks. This procedure provided a very accurate pose-estimation, (average error around 1 cm), when compared with that measured by hand. It was estimated that we can measure the real position by hand with an accuracy of approx. 1 cm. Figure 4.30 represents the automatic initialization procedure based on three green landmarks. In particular, the figure shows the simple setup consisting of three green papers hanging on the wall at known positions (right-hand), and a sketch of the triangulation procedure (left-hand).

Nineteen runs were performed. In all runs three independent processes were running simultaneously. The first process updated the robot position by only reading the odometric values. The second one used the "angular deviation" to a single landmark method, (subsection 3.1.2), and the third one used the proposed triangulation of three landmarks. The system "got lost" in three runs, (due to a large odometric errors), while in only one run we had a final error above the 10 cm. In the other fifteen runs the system succeeded with a final error up to 5 cm, in average 3 cm.

Figure 4.31 shows examples of typical optimal triplets which were selected during the tested runs in relation to different robot positions. Figure 4.32 shows the output of a typical run representing the X, Y position of the computed robot path. Figure 4.33 shows, for the same run, the measured error at *check-points*, that is, those points where the estimated position was compared with the real position (measured by hand). The diagram shows that the error of the alternative method inexorably increases, (as with odometry), meanwhile triangulation keeps it small.

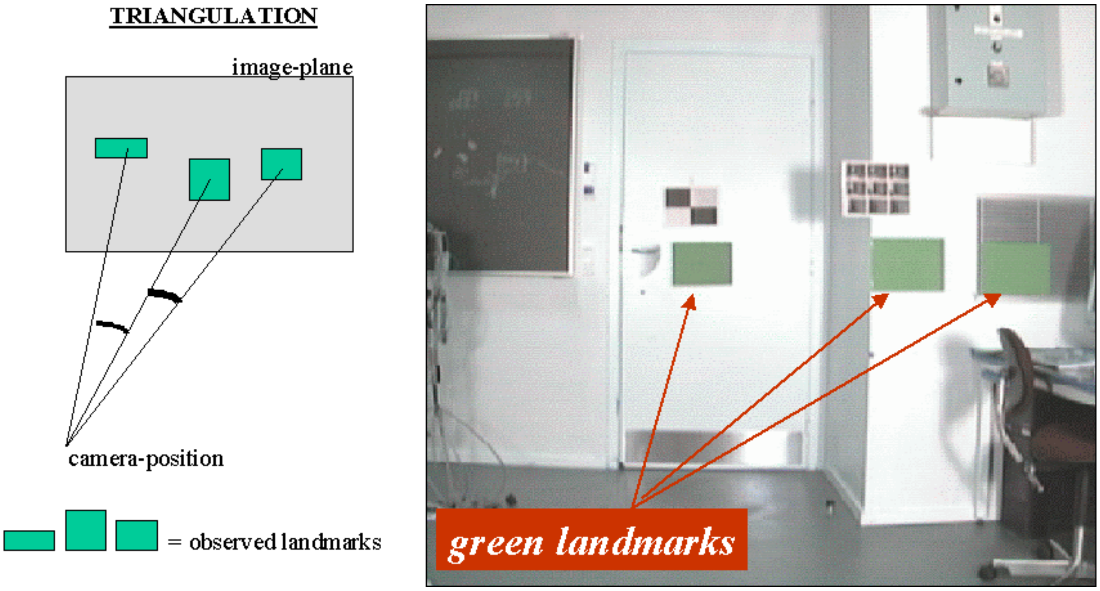


Figure 4.30: Automatic initialization procedure based on three green landmarks. The figure shows the simple setup consisting of three green papers hanging on the wall at known positions (right-hand), and a sketch of the triangulation procedure (left-hand).

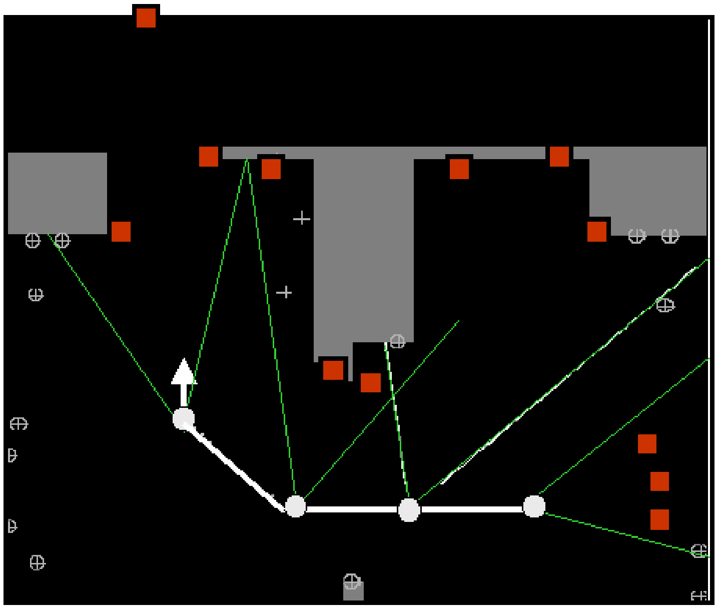


Figure 4.31: The figure shows a typical path on the laboratory floor-map. The given path represented a trajectory that robot had to run along in order to get the laboratory exit-door. The green lines emanating from robot positions (white circles) represent the camera field of view. The little red squares represent selected landmarks.

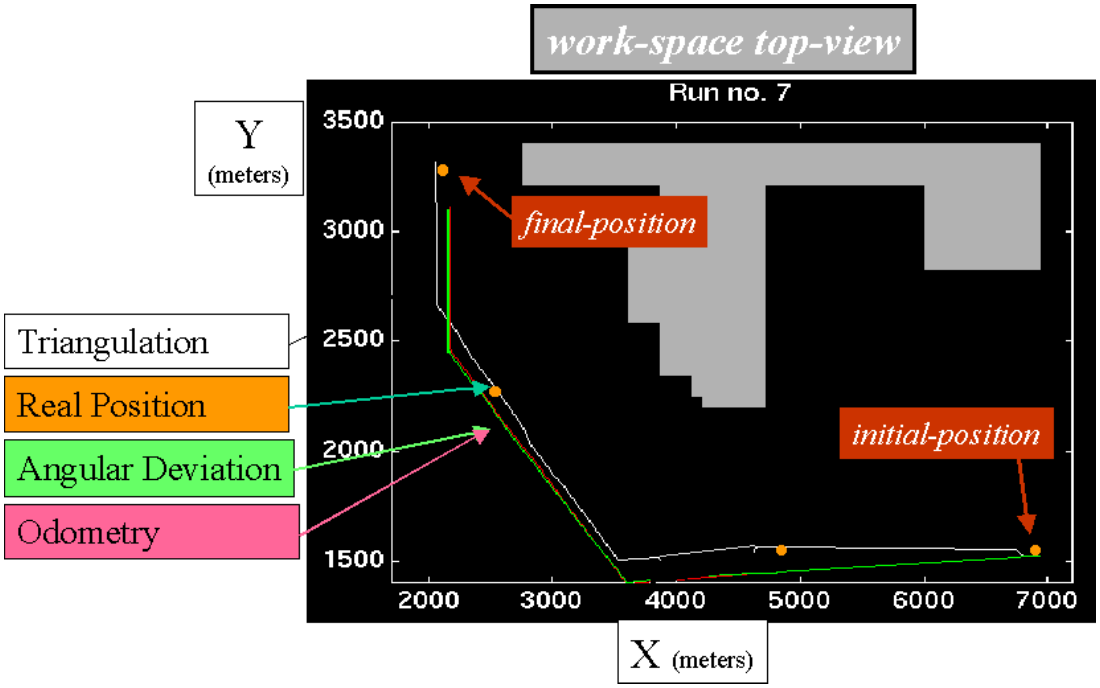


Figure 4.32: Diagrams of a typical run showing the X, Y position of the computed robot path. The white line represents position computed by the triangulation process. The pink line represents position computed by the odometric and the green line the angular deviation to a single landmark processes. Circles represent the real positions.

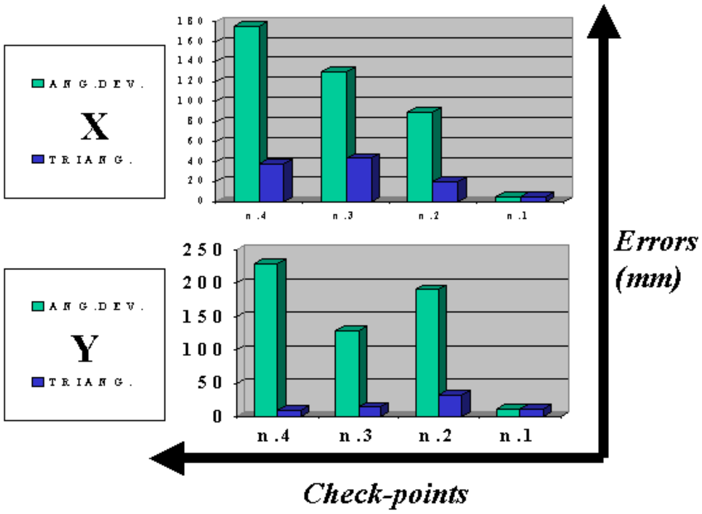


Figure 4.33: Errors in the estimated robot position compared with the real position (measured by hand). In green positions estimated using the "angular deviation to a landmark" method. In blue positions estimated using the proposed triangulation method.

The robustness of proposed triangulation method was also tested. Figure 4.34 top-diagram shows the case when an error in the localization algorithm, arising during one run at a certain position (point A), does not affect on the next pose estimation (point B).

It was also interesting to test how the three tested localization processes, (odometry only, angular deviation, triangulation), react to a large initial error. This led us to perform some extra runs where errors were introduced to provided initial position. The bottom diagram in figure 4.34 shows the output of one of these runs. The system was able to recover a positional error up to 8 cm in the gaze direction of the camera and 50 cm in the direction perpendicular to the gaze. This result shows that an accurate initial estimate it is not needed.

It is possible to consequently conclude that in most cases of unexpected positional errors, the triangulation method is able to keep track of the correct position whereas the alternative methods fail. In fact, the angular deviation to a single landmark method can not recover an unexpected error.

The results achieved from performed experimentation were very encouraging because they demonstrated that the accuracy and robustness of proposed localization strategy based on optimal-triplet selection and triangulation was effective in a real setting, being the robot able to accurately estimate robot position and do not accumulate error.

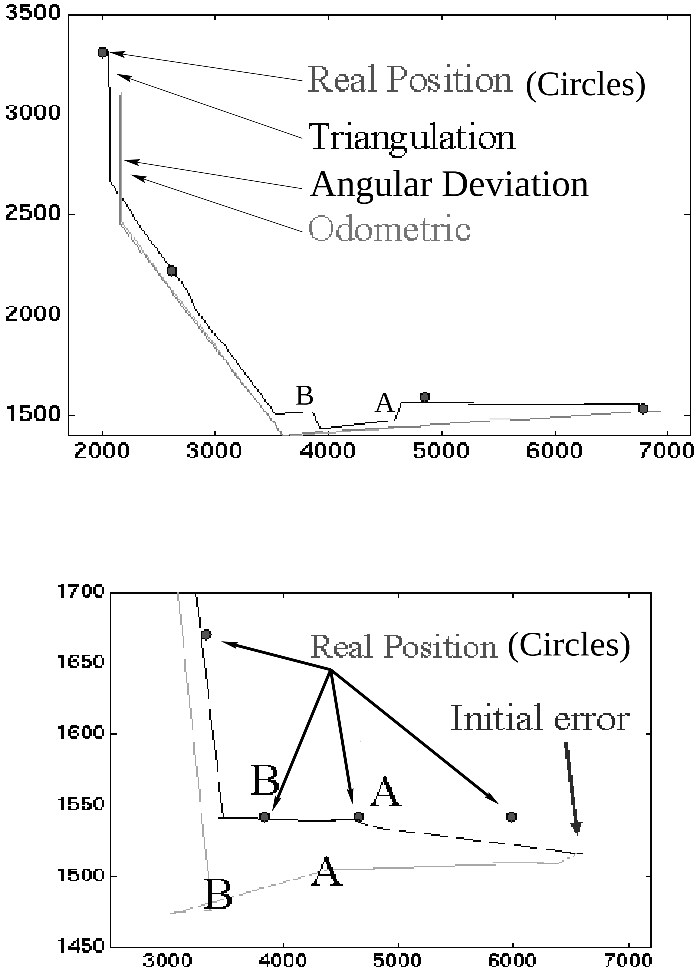


Figure 4.34: Diagrams showing fault-tolerance of the proposed method while robot is following an assigned path. The black line represents position computed by the triangulation process, while the two almost overlapping grey lines represent position computed by the odometric and the angular deviation to a single landmark processes. The top diagram shows the case when an error in the triangulation-based localization algorithm, arising during one run at a certain position (A), does not affect on the next pose estimation (B). The second diagram shows a case where an error in the initial robot position was introduced. The triangulation method is able to keep track of the correct position whereas the alternative method fails.

4.5 Summary

In this chapter the proposed method for automatic recognition of visual landmarks manually learned has been described and its experimentation presented, (section 4.1). The proposed scheme represent an accurate and reliable solution model as well as the basis for a more challenging solution approach, (chapter 6), involving visual landmarks automatically learned. The experimental results were very encouraging because they demonstrated the advantage of proposed recognition method based on reprojection of reference-views, which can be quantified in terms of a "wide" observation-area, (i.e. "wide" recognition possibility).

The proposed method for robot pose computation based on triangulation and optimal triplet selection has also been described (section 4.2), together with a brief analysis of main problems, mainly related to uncertainties in the input parameters which propagate to output response. The results achieved confirmed the potential of proposed optimal-triplet selection and triangulation. The system was in fact able to estimate robot position with sufficient accuracy and do not accumulate error. Few guidelines about how to approach the problem of propagating uncertainties have also been discussed and it seems that a lot can be done with future research activities, and discussion in this chapter paves the way towards it.

Still, as came out from presented experimentation, and as seen in the literature, it is not possible to achieve a very precise positional information for a mobile robot. Nevertheless, the proposed localization system aims to decrease the amount of human assistance required by a robot to navigate indoor environments by a more precise pose estimate and a recoverable positional error.

The key issues characterizing the proposed self-positioning method are:

1. a landmark triplet selection technique which minimizes expected positional error (optimal triplet selection);
2. an accurate detection of landmark location in the image-plane of current view (visual landmark recognition);
3. an effective exploitation and combination of multiple environment "features" (triangulation).

The approach proposed in this chapter for robot self-localization represents the basis to the development of the self-relying localization system proposed in this thesis, hence, as discussed in chapter 3, the first researching step for the proposed method.

In fact, it is based on the self-localization method presented in this chapter, and on the achieved experimental results, that the other key system components, (such as landmark view acquisition, landmark pose computation, realistic synthesis of landmark appearance, automatic recognition of self-learned landmarks, etc.), are designed.

Consequently, it was important to first make sure that the potential of proposed visual landmark recognition, optimal-triplet selection, and triangulation method, was effective in a real system setting, in order to propose an entire autonomous robotic system relying on this type of self-localization strategy.

Chapter 5

Automatic Learning

Fully autonomous navigation requires the capability of automatically learning the information required for robot self-localization. This means an automatic acquisition of the environment model which in our case is composed by landmark representative images and their position in the environment.

Typically, for autonomous systems proposed in the literature, there is a learning phase prior to a localization phase, [152], [41], [9], [141]. Following this guideline, an automatic learning is proposed to our system as first action to be taken by the robot. The learning procedure can also run later during robot navigation, every time the system requires to extend or improve its current knowledge of the workspace, or in case the robot positional error becomes too large.

As described in sub-section 3.1.5, the method for automatic learning must be able to: (1) automatically acquire landmark representative images; (2) automatically estimate landmark pose in the environment; (3) allow for "accurate" visual and positional information.

The problems of computing robot pose and learning landmark visual and positional information are coupled, since the robot localization performance is strongly depending on the proposed strategies for learning the environment. In particular, the landmarks acquired in the learning phase have to be recognized during navigation in the incoming images. Recognition of previously acquired landmarks is a complex and difficult task which strongly depends on the fidelity of acquired landmark views and on the accuracy of estimated landmark poses.

How visual selection and pose computation can be performed and how this can help selection of suitable landmarks for localization, gives rise to a number of questions. For example, what is useful information? what is an appropriate selection criteria? what is a convenient robot configuration for pose learning? Some of the above mentioned problems have been addressed in this chapter and a solution is proposed.

At the initial stage the robotic system does not posses any priori knowledge of the workspace except for the internal camera parameters. It is consequently based on this knowledge that the system has to be able to learn the environment model. In particular, the proposed method for model learning requires the robot to periodically stop at arbitrary locations, and while staying at these locations to run a learning procedure in order to acquire landmark representative views and their associated poses.

The robotic system might be provided with its initial position. Though, the initial robot position is not essential information, and the position assumed by the robot at the time zero could be considered the origin of the global coordinate system. Nevertheless, if a workspace map is intended to be provided to the system as prior knowledge, a global coordinate system should be set, and the initial robot position should be provided to the robot at time zero. The initial position could be provided either by hand or learned through a special setup. As for the latter a solution based on triangulation of green landmarks, (the technique presented by the author in [93]), has been proposed and thoroughly tested, (section 4.4).

The general assumption for the learning phase is that the robot knows its pose with low uncertainty. This in fact is what is supposed to happen during robot navigation when the system runs the localization procedure presented in chapter 4. As for the initialization time, i.e. when robot positional information may not be available, low positional uncertainty is supposed to be achieved by letting the robot run along simple and short range (less than 5 meters) trajectories, which in principle do not allow the odometric system to accumulate large errors (odometric readings can be trusted). This issue is discussed further in section 5.4.

The following sections will thoroughly look into the main method functionalities for what concern the proposed learning strategy. The first section deals with the problem of acquisition of landmark candidate views, starting with problem analysis, (sub-section 5.1.1), and then presenting the proposed method for workspace-views acquisition, (sub-section 5.1.2), and landmark view extraction (sub-section 5.1.3). Experiments are described at the end of the section (sub-section 5.1.4).

The second section deals with the problem of estimating landmark position and orientation, starting with presenting the core idea, (sub-section 5.2.1), and then proceeding with the proposed solution in terms of correspondences test, (sub-section 5.2.2), and pose extraction, (sub-section 5.2.3). Experiments are described at the end of the section (sub-section 5.2.4).

The third section analyzes and characterizes the learned information so that only the most reliable information (the *elected* landmarks) is stored as the learned environment model, (sub-sections 5.3.1 and 5.3.2). The navigation strategy related to presented learning method is then discussed in section 5.4. Eventually, a summary of the entire learning process (and learned lesson) is provided, (section 5.5).

5.1 Landmark View Acquisition

The acquisition of landmark visual appearances is a necessary learning step for our system. The proposed method must in fact allow for extraction of high fidelity and discriminant environment visual-appearances to be later used as reference landmark views. Furthermore, the extraction process has to operate fully autonomously. A method is then proposed to: (1) acquire workspace views; this computational step is called: *Panoramic View Synthesis* (sub-section 5.1.2); (2) detect on the acquired views image regions which are suitable for being landmarks; this computational step is called *Landmark Candidate Extraction* (sub-section 5.1.3).

5.1.1 Problem Analysis

Visual landmarks are characteristics of the environment which are observed by a camera and this leads to the fact that they may appear really different during learning and localization. The appearance of a landmark into an image-plane depends on many aspects. They mainly are: landmark geometry, (i.e geometry of represented environment characteristics), landmark size and texture, environment light-condition, distance and angle between camera and landmark.

When landmarks represent stable characteristics of the environment, the landmark represented geometry and its texture-structure are not expected to change once we have chosen our landmark. On the contrary, environment light-condition, distance and angle to a landmark may be different between learning and localization time, and the system performance is likely to be affected by such changes.

The appearance of a landmark in the image-plane is very sensitive to changes with environment illumination-conditions, since this could lead to undesired reflections, highlights and shadows. The appearance of a landmark is very sensitive to changes in distance and angle to a landmark, since different aspects of objects may become visible, and occlusions and perspective distortions may be generated.

The appearance of a landmark in the image plane also is very sensitive to errors in landmark and robot pose, leading to the fact that the appearance of a landmark into the image-plane is not as expected.

In particular, the error in landmark pose arises during learning and it is mainly due to algorithmic errors (numerical approximations) or inaccuracies in previous robot pose estimation, e.g. an approximated initialization. The error in robot pose arises during learning and localization and it is mainly due to errors in landmark pose, odometric estimate, or changes of landmark texture in the camera image-plane.

The changes of landmark textures may be originated by robot or object movements or by people moving around in the environment. The effect of such changes may cause different aspects of objects to become visible, and occlusions and perspective distortions to be generated.

The errors in landmarks and robot pose may also lead to an error in the location of landmarks in the image plane. This error may arise from the fact that landmarks may be confused (e.g. because they become very similar), or worse misidentified (e.g. because objects may become similar to target landmarks when partially occluded).

All these missed or wrong identifications can produce errors in robot pose estimation, and can make the navigation process unreliable or unstable. Figure 5.1 summarizes the main aspects affecting the landmark appearance in the image plane.

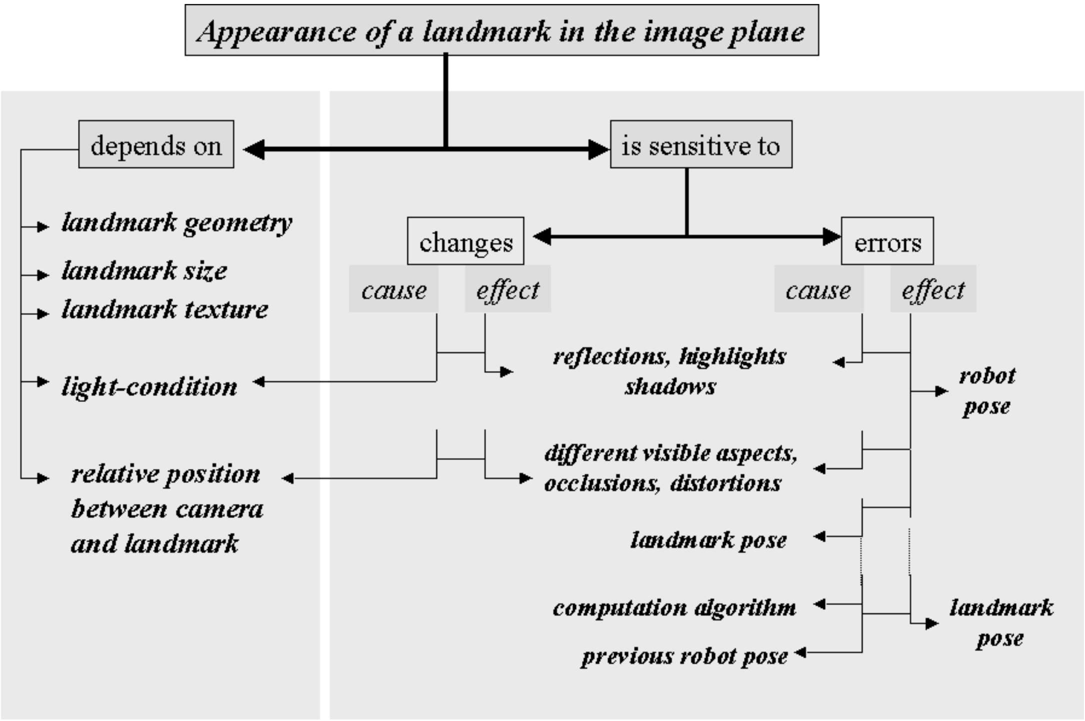


Figure 5.1: The figure represents the main aspects affecting the landmark appearance in the image plane. In particular, the figure shows the aspects on which the appearance of a landmark *depends on*, and the aspects on which the appearance in the image-plane *is sensitive to*. It can be seen that the effects of some changes may become cause of error. For example changing in light-conditions may generate reflections, which can lead to errors in robot estimate position during localization.

A main challenge for the learning phase consequently is selection of reliable landmarks. In particular, the selected landmarks have to be robust to changes in environment conditions, camera position, and errors affecting camera position.

It has been necessary to make some assumptions in order to reduce the complexity of the problem: the environment is supposed to be mostly static, that is, only a few changes may occur, and illumination can vary moderately. In addition, it is chosen that landmarks should be characteristic of the environment which represent planar surfaces. Hence, a method is proposed in the next pages in order to automatically learn landmark representative images and to accurately estimate landmark poses.

5.1.2 Panoramic View Synthesis

The first action to be taken by the vision system is acquisition of workspace images, since from these images landmark representative views must be extracted.

The same vision system has to be used during both the phase of learning and the phase of localization. Consequently, it must be designed in order to match requirement of both the phases. In particular, during the learning phase the vision system must allow extraction of landmark representative images according to the required landmark-image definition, (see sub-section 3.1.3), and during the localization phase the vision system must allow for recognition of the acquired landmarks and the observation of at least three landmarks in the same camera image plane, (chapter 4).

When using a standard CCD camera the main camera parameters to be set are focal-length and view-field. But, what is an optimal condition for image acquisition? and how can we acquire representative views?

The optimal condition for image acquisition was discussed in sub-section 4.1.2, where it is stated that in case of a texture representing a planar surface, the most representative view is a frontal view, and that the ideal acquisition situation would then be to capture occlusion-free frontal-views of landmarks, taken at a convenient distance. (such as shown in figure 5.2).

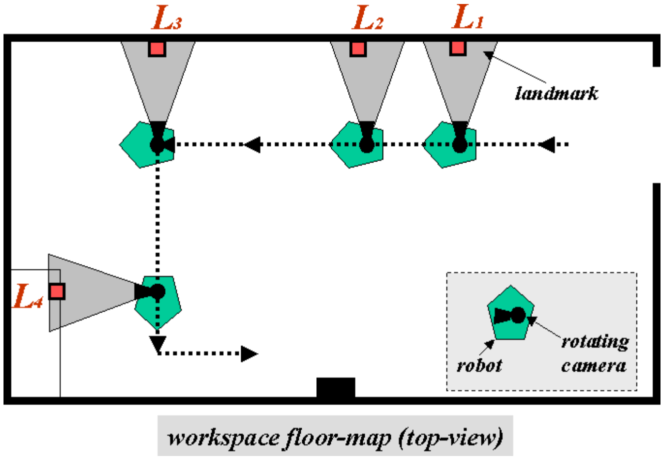


Figure 5.2: The figure represents the ideal situation setup for proposed *manual training*.

Unfortunately, to reproduce such ideal situation is not practical and difficult to make automatic. The robot would in fact need a certain level of prior knowledge about the environment and feature locations, in order to reach the ideal robot-feature configurations. To posses such a knowledge is against the principle of a fully autonomous navigation, (section 1.3).

Since the robot instead possesses no knowledge about its position related to environment features, it is likely to happen that a current robot position is not ideal for observation of certain environment features. The same position could however be suitable for observation of some others environment features, which leads to the conclusion that having a wider camera field of view will increase the probability of containing representative views of some workspace areas in a single image.

A wide camera field of view represents an advantage since it simplifies the acquisition process and subsequent landmark extraction. Unfortunately, a wider field of view typically means a lower image definition, and there is a limit which can not be passed. The limit depends on the required image definition that has to be guaranteed.

As example, the use of a wide angle camera, such as the commercially available omnidirectional cameras, would be very useful, but this is unfortunately not acceptable. The images obtained have substantial distortions and mapping a wide scene into the limited resolution of a video camera compromises image quality. The provided image definition will thus be too low for the proposed method, (it does not satisfy the requirements discussed in sub-section 3.1.3).

Photo Mosaicing

The proposed solution which match both the requirements of wide field of view and high resolution, is represented by a system able to collect high resolution images of the workspace in a practical way, and then fuse them into a single image.

A common solution in the literature is *photo-mosaicing* which consists on merging an image sequence taken from different point of views in order to obtain a more complete view. The major issues in photo-mosaicing are aligning, pasting, and blending, frames in a video-sequence, to enable a more complete view. In particular, image-alignment determines the transformation that aligns the images to be combined into a mosaic, and usually uses rigid transformations as picture translation and rotations; image-pasting identifies regions in different images which are overlapping, and then determines some kind of a combination of all overlapping images; and image-blending determines how to overcome the intensity difference between images.

In all cases images are aligned pairwise, using a parametric transformation like an affine transformation or planar-projective transformation. A reference frame is selected, and all images are aligned with this reference frame and combined to create panoramic mosaic.

Aligning all frames to a single reference frame is reasonable when the camera is far away from observed objects and its motion is mainly a translation and a rotation around the optical axis. Significant distortions are on the other hand created when camera motions included other rotations, or when the camera is relatively close to observed objects.

The technique of Manifold Projection[1], (Peleg and Herman [114]), overcomes many difficulties in photo-mosaicing under the assumptions of insignificant change of scale and limited parallax. This enables a fast creation of low-distortion panoramic mosaicing under general camera conditions. In our case, however, if we assure well controlled movements of rotation, a simpler mosaic technique can successfully be applied to images representing objects at relatively close distances, (as happen in typical domestic indoor-environments).

The proposed mosaics is generated by rotating the camera around its optical center using a special but simple device, i.e a pan unit able to freely rotate on 360 degrees, and then creating a panoramic image which represents the projection of the scene onto a cylinder,

[1] The basic principle is the alignment of the strips which contribute to the mosaic, rather than the alignment of the entire overlap between frames. Each strip undergoes the minimal distortion to be aligned with the neighboring strip, eliminating global distortions.

(Yuen and MacDonald [150], Benosman and Kang [11]). In particular, it is proposed that the workspace views are collected by panning the camera around the vertical axis passing through its optical center, and the panoramic views are synthesized by exploiting cylindrical projection as in McMillan and Bishop [103].

Synthesis by Cylindrical Projection

The algorithm for synthesizing panoramic-views is based on the principle that any two planar perspective projections of a scene which share a common viewpoint but with different viewing directions are related by a two-dimensional homogeneous transform H, as showed in the following equation, where x and y represent the pixel coordinates of an image I, and x' and y' are their corresponding coordinates in a second image I'.

$$\begin{bmatrix} u \\ v \\ q \end{bmatrix} = H \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{5.1}$$

$$x' = \frac{u}{q} \qquad y' = \frac{v}{q} \tag{5.2}$$

This well known result has been reported by several authors, (e.g. Heckbert [69]). The images resulting from typical camera motions such as pan, (but also tilt, roll and zoom), can all be related in this fashion. When creating a cylindrical projection, we will only need to consider panning camera motions.

In order to project an individual image onto a cylinder, we must first determinate a model for the camera projection or, equivalently, the appropriate homogeneous transforms.

Many different techniques have been developed in the literature for inferring the homogeneous transform. For example, Heckbert [69] involves establishing four corresponding points across each image pair, Mann and Picard [98] and Szeliski [136] derive the homogeneous transformations without specific point correspondences, etc. However, a very precise estimation of camera internal parameters and panning angle is important in our case, because landmarks might represent relatively large image regions. Consequently, it is suggested that the determination of camera internal parameters is precisely estimated off-line, and the panning angle is provided by the hardware.

The determination of the homogeneous transformation is consequently straight forward by applying equation 5.4. In particular, the homogeneous transformation can be decomposed into two parts which include a projection matrix, P, which is determined entirely by camera properties, and an extrinsic transform, Rot, which is determined by the rotation around the camera center of projection. In equation 5.3, γ is the panning angle, f is the camera focal length in pixels, C_x and C_y are the pixels coordinates of the intersection of the optical axis with the image plane, η is a skew parameter representing the deviation of the sampling grid from a rectilinear grid and ρ is the sampling grid's aspect ratio.

$$Rot = \begin{bmatrix} cos\gamma & 0 & sin\gamma \\ 0 & 1 & 0 \\ -sin\gamma & 0 & cos\gamma \end{bmatrix} \qquad P = \begin{bmatrix} 1 & \eta & -C_x \\ 0 & \rho & -C_y \\ 0 & 0 & f \end{bmatrix} \tag{5.3}$$

$$H = P^{-1}\ Rot\ P \tag{5.4}$$

Due to the quality of modern frame-grabber hardware, skewing can be disregarded ($\eta = 0$).

Using the above equations and knowing the contained parameter values, it is then possible to compose mapping functions from any image in the sequence to any other image as described by the following equation:

$$H_i = P^{-1}\ Rot_1\ Rot_2\ Rot_3\ Rot_4\ \cdot\cdot Rot_i\ P \tag{5.5}$$

Since each image pixel determines the equation of a ray from the center-of-projection, the re-projection process merely involves intersecting these rays with the projection manifold (in this case a cylinder). Figure 5.3 illustrates the method.

Figure 5.4 shows examples of synthesized panoramic views. Each panorama is generated from a dense sampling of workspace views taken from the same location but letting the camera panning over 60° (top-row) and 360° (middle-row), relative to the robot heading. In particular, figure 5.4 bottom-row shows a panoramic view over 360° related to a different workspace. In case of the 360° panoramas, 72 images were acquired for each location, (i.e. one image each 5 degrees rotation). More examples are discussed in the experimentation subsection, (see 5.1.4), where also details about the experimentation setup are provided.

The proposed solution represents a compact and practical solution, which requires a simple setup and fit the available hardware. A significant advantage of a cylindrical projection is the simplicity of the acquisition: a video camera and a pan unit capable of continuous panning.

The synthesized panoramas represent high-resolution workspace-views (they are based on many high-resolution views). In particular, they allow observation of wide workspace-areas and provides the system with sufficient image definition. For these reasons the proposed solution has been preferred to alternatives.

The drawback of the proposed solution is represented by the time required by the hardware to collect the views. The robot typically needs to stop during a learning phase every time a new panoramic view is required.

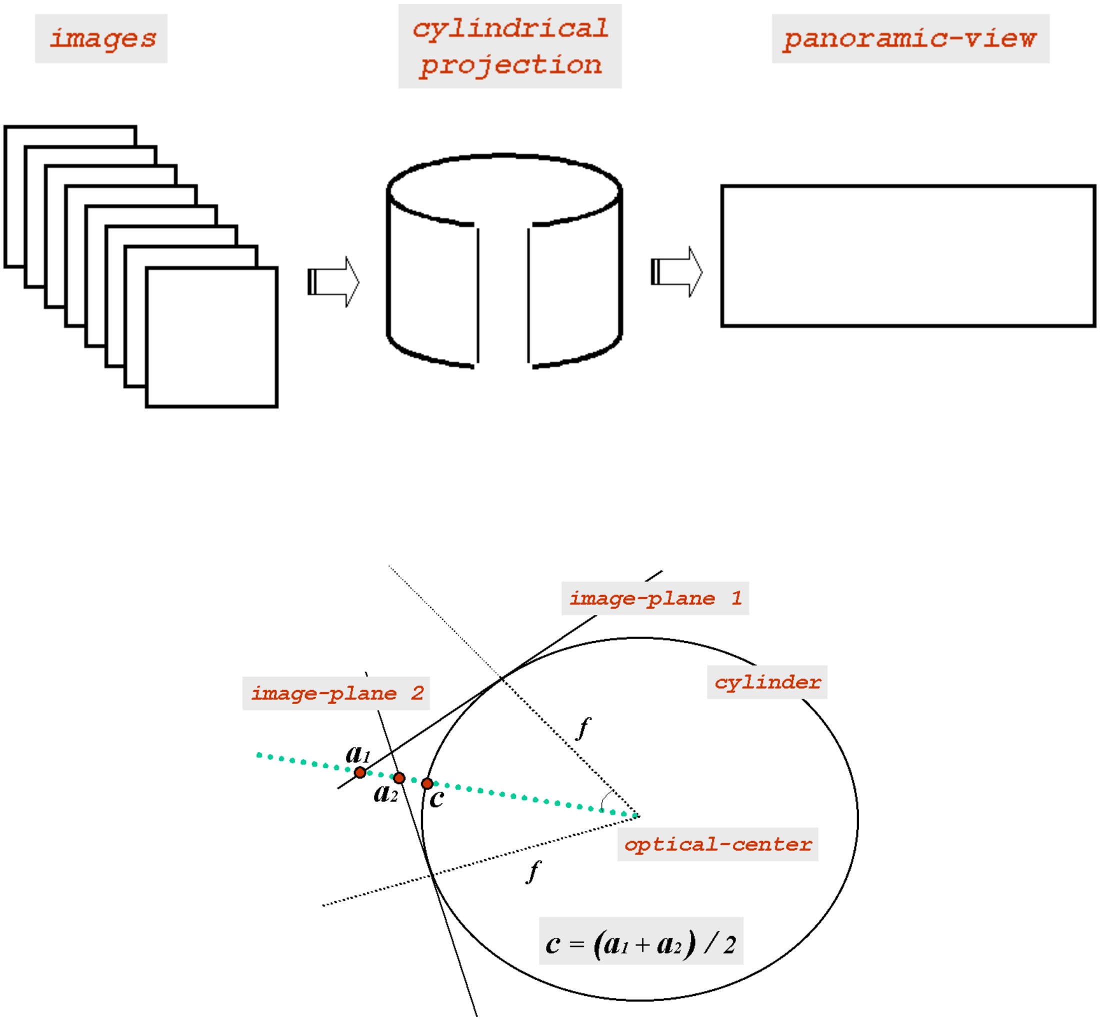


Figure 5.3: Panoramic view synthesis based on cylindrical projections. The figure top-row illustrates the flow of information going through main "data structures", and the figure bottom-row illustrates an example of what could be the simplest way to calculate texture information in the cylindrical view. The variables a_1, a_2, and c represent pixel values.

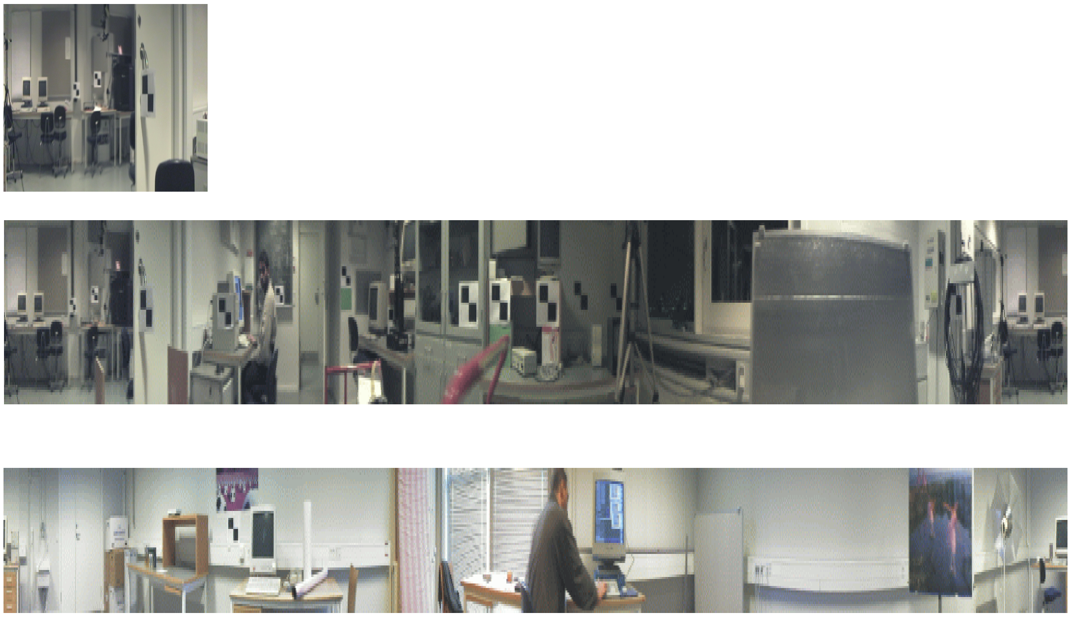


Figure 5.4: The figure top-row shows a computed panoramic view which has been synthesized by projecting 10 images of 350x350 pixels each, over 60 degrees, (i.e. 1 image every 10 degrees rotation). The figure middle-row shows a panoramic view of the same workspace but synthesized projecting 72 images of 350x350 pixels each, over 360 degrees, (i.e. 1 image every 5 degrees rotation). The overlapping area between two adjacent images is in this case about 85%. Maximum overlap is 7 images contributing to the same pixel in the panorama. The figure bottom-row shows a panoramic view of a different workspace, which is still synthesized by projecting 72 images over 360 degrees, but the size of each image is 256x256 pixels. The entire image was 256x4101 pixels.

5.1.3 Landmark Candidate Extraction

Fully autonomous navigation implies the use of landmarks representing structures already in the environment. In particular, we look for landmarks representing discriminant textures. The characteristic for a landmark of being discriminant is referred as to have a *high-contrasted* and *unambiguous* texture.

Since the landmark detection process is vision based, (i.e. it is in the camera image plane that landmarks have to be identified), the algorithm used to extract texture patches has to handle the large amount of data contained in an image and reducing it in an advantageous way.

But, what is useful information? what is an appropriate selection criteria?

The problems mentioned above have been addressed in previous research works. It is thus opportune to this work to rely on previous experiences for attention strategies and so design an attentive scheme based on the learned lesson.

Attention Selection Mechanism

From empirical derived knowledge and the analysis presented in Andersen [2], it has been chosen to use the method invented by Culhane and Tsotsos, [37], as the basis for the attentional mechanism. This in order to produce as output a set of texture patches which may be suitable as landmark visual template.

The basic structure of the Culhane-Tsotsos algorithm is described in the 5 steps of the algorithm described following and illustrated in figure 5.5. The technique builds a hierarchy of representations which input layer is represented by the camera image and relies on the assumption that the value of cells in the input layer directly reflects how salient a specific location is. Thus, for example, in case where an intensity image is used as input, the technique will select bright areas, (i.e. high pixel intensity values).

The hierarchy is built using averaging of the pixels in the layer below. The most interesting region is then selected starting at top layer in the hierarchy. Culhane and Tsotsos introduce the notion of receptive fields which correspond to a kind of region of interest. These receptive fields, of varying size and shape compete in a winner-take-all search strategy, and the receptive field with maximum saliency wins. The pixels in the layer below the current which are connected to the pixels within the winning receptive field, are then selected for processing. Then, the entire process is repeated. When the input layer is reached, selection of the most salient area has been achieved.

When the processing is done, the selected area and all the neighboring pixels which competed in the winner-take-all are inhibited. This is done by setting the saliency value to zero. The selection is thus completed and a new selection can start. This is performed by issuing a total re-computation from ”step 2” of the algorithm following described. The process is run until some criterion is meet, or until no salient features are present in the input map. Figure 5.5 shows a visual representation of the Culhane-Tsotsos proposed technique.

```
1. Receive input from feature extraction process at lowest level
Repeat
    2. Build a hierarchy using weighted sums of previous layer
    Go to top layer (lowest resolution)
    Repeat
        3. Run Winner Takes All at current level
        Go to next layer (higher resolution)
    Until bottom layer is reached and processed
    4. Process data in selection region
    5. Inhibit data in selection and inhibition region of input layer
Until no more regions can be found
```

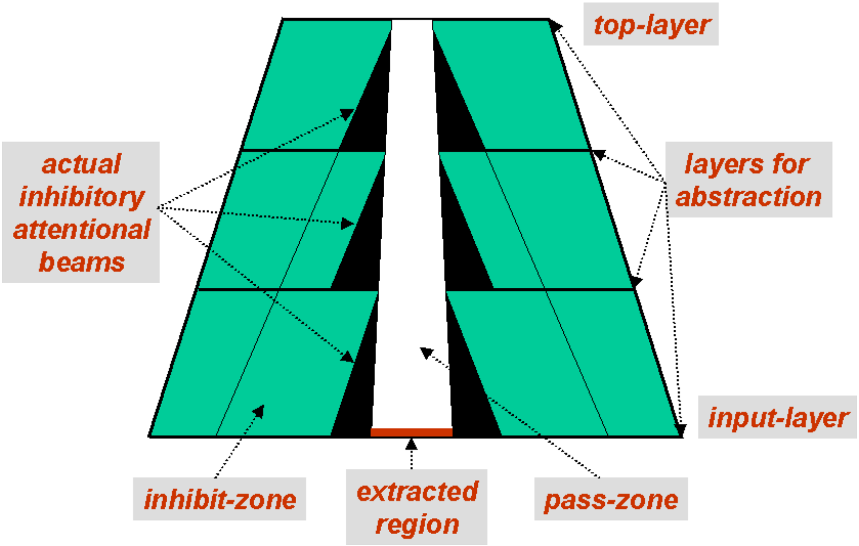


Figure 5.5: The figure represents the 1D visual presentation of the attention selection mechanism proposed by Culhane and Tsotsos, [37]. The beam "shines" from the top of the hierarchy towards to the input layer, illuminating the area of interest, [2].

Culhane and Tsotsos emphasize that automatic selection of receptive fields size is possible. They argue for a biological inspiration approach where the receptive fields of different size compete for highest saliency. There is however, as experienced, some tuning involved depending on the application, but certainly their work has resulted in some promising ideas.

The number of receptive fields which needs to be considered for each location in a given layer is computed by the equation 5.6, where min and N denotes the minimal and maximal side length of the receptive fields in the two directions. Thus the equation counts the total number of different receptive fields that may be created within the largest receptive field N_x x N_y.

$$N_{RF} = \sum_{k_y=min_y}^{N_y} \sum_{k_x=min_x}^{N_x} (N_y - k_y + 1)(N_x - k_x + 1) \tag{5.6}$$

In order to limit the number of receptive fields to consider, they proposed to use rectangular receptive fields bounded by established maximal and minimal side lengths. For the proposed work the range of 28x28 to 36x36 was set as ideal side lengths for the extracted landmarks. This based on the fact that this range was successfully experimented in the self-localization algorithm, (described in chapter 4). In our case, thus, min_x and min_y are set to 28 and N_x and N_y are set to 36. The computation of each receptive field is very simple since it is a weighted average of the pixels it covers.

Feature and Conspicuity Maps

Generally, more elaborated input images may be provided to the attention mechanism through a saliency map. This also includes incorporating multiple features, which was not covered in the original attention selection algorithm.

But, what type of image-feature should a landmark contain in order to be discriminant?

The corner feature is considered a useful feature to be included to the computation of the saliency map. It represents image regions high-contrasted and invariant to camera viewpoint. This allows for a easier detection by a specific filter mask. The characteristic of being an invariant feature also implies being suitable for re-projection. Unfortunately, since corner features are abundant in typical indoor scenarios, they can easily be confused with each other, and thus misidentified. Nevertheless, this drawback can be reduced by integrating corners with other features.

Corners are extracted by a method proposed by Zhang et al, [151], which represents a Canny filter followed by a covariance analysis. The idea is shown in equation 5.7, where I_x and I_y are the magnitudes of directional derivatives. The parameter k is a constant which may be varied. The authors argue for a value of 0.04, which has been found appropriate here as well.

$$c(x, y) = I_x^2 + I_y^2 - 2I_xI_y - k(I_x^2 + I_y^2)^2 \tag{5.7}$$

A corner is detected when the result of equation 5.7 is above some given positive threshold, which needs to be set according to the scenario. I_x and I_y may be found using the canny filter. Figures 5.8 and 5.9 second-rows from the bottom, show detected corners. More examples are shown in the experimentation section, (5.1.4).

The edge feature posses the same characteristics of the corner. However, only vertical edges are considered, since the proposed vision system has a camera rotating around a vertical axis. The horizontal edges would not well constrain the rotation of the camera around the vertical axis (or position).

Edges are extracted using high pass filtering. In particular, it was used the recursive Canny - Deriche filter, [45]. This filter is equivalent to the first order derivative. Figures 5.8 and 5.9 second-rows from the bottom show detected edges. More examples are shown in the experimentation section, (5.1.4).

As we can see, the characteristic for an image region of being high-contrasted can be obtained by using specific filters allowing detection of geometric 2D features. However, common geometric features may easily become ambiguous.

The ambiguity problem can be reduced by selecting image regions where number and contrast of a certain feature is higher. This can be done by using the Difference of Gaussian (DoG) filter. Examples of DoG performance for different image-features can be found in figures 5.8 and 5.9 third-rows from the bottom. In all images the variances has been kept fixed for the two Gaussian's. More examples are shown in the experimentation sub-section, (5.1.4). As it can be seen from the figure, the DoG filter acts like a high pass filter (second order derivative). These regions with non varying intensity in feature responses are suppressed, while regions where features are located in small clusters are amplified.

The local symmetry is another feature included to the computation of the saliency map. It represents image regions containing symmetric patterns in local neighborhoods which can be described by a set of functions. The local symmetry detector was designed to search for similarity along circular and straight lines, getting inspiration from what described in Hansen, [64], and in Granlund and Knutsson, [60]. Figure 5.6 shows examples of local symmetry patterns which were detected by the proposed detector. The patterns which are represented in the figure could be weighted differently in our detector depending on the type of texture expected in the environment. A local symmetry typically identifies circular regions, making it suitable to fit squares or rectangular regions, i.e. such as the required texture-patch shapes.

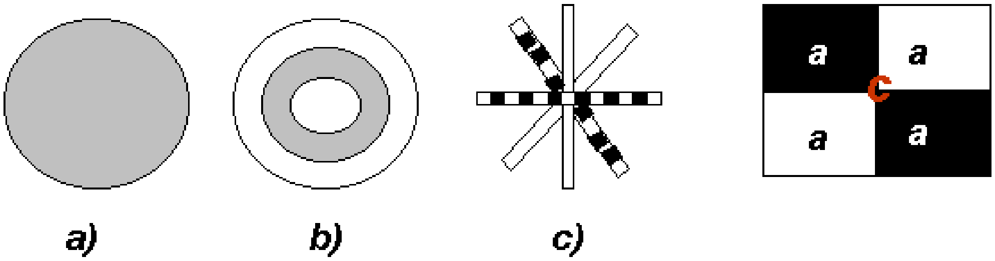


Figure 5.6: The figure shows examples of local symmetry patterns which were detected by the proposed algorithm. The patters which are represented in the figure could differently be weighted in our detector depending on the type of texture expected in the environment.

Symmetric patterns do not necessarily represent high-contrasted areas depending on which type of symmetry is represented. Nevertheless, when a Difference of Gaussian filter runs on the output of a local symmetry detector, it generates a conspicuty map containing multiple local symmetries which may represent high contrasted regions. Figure 5.7 shows this case. In particular, images at the right-hand side show examples of regions selected after that the proposed attention selection mechanism was run on the feature map (top-row) and on the feature-map filtered by the DoG, i.e. the conspicuity map, (bottom-row). The latter generates more discriminant selected patches and it shows how the ambiguity problem can be reduced by the Difference of Gaussian filter also in case of local-symmetries.

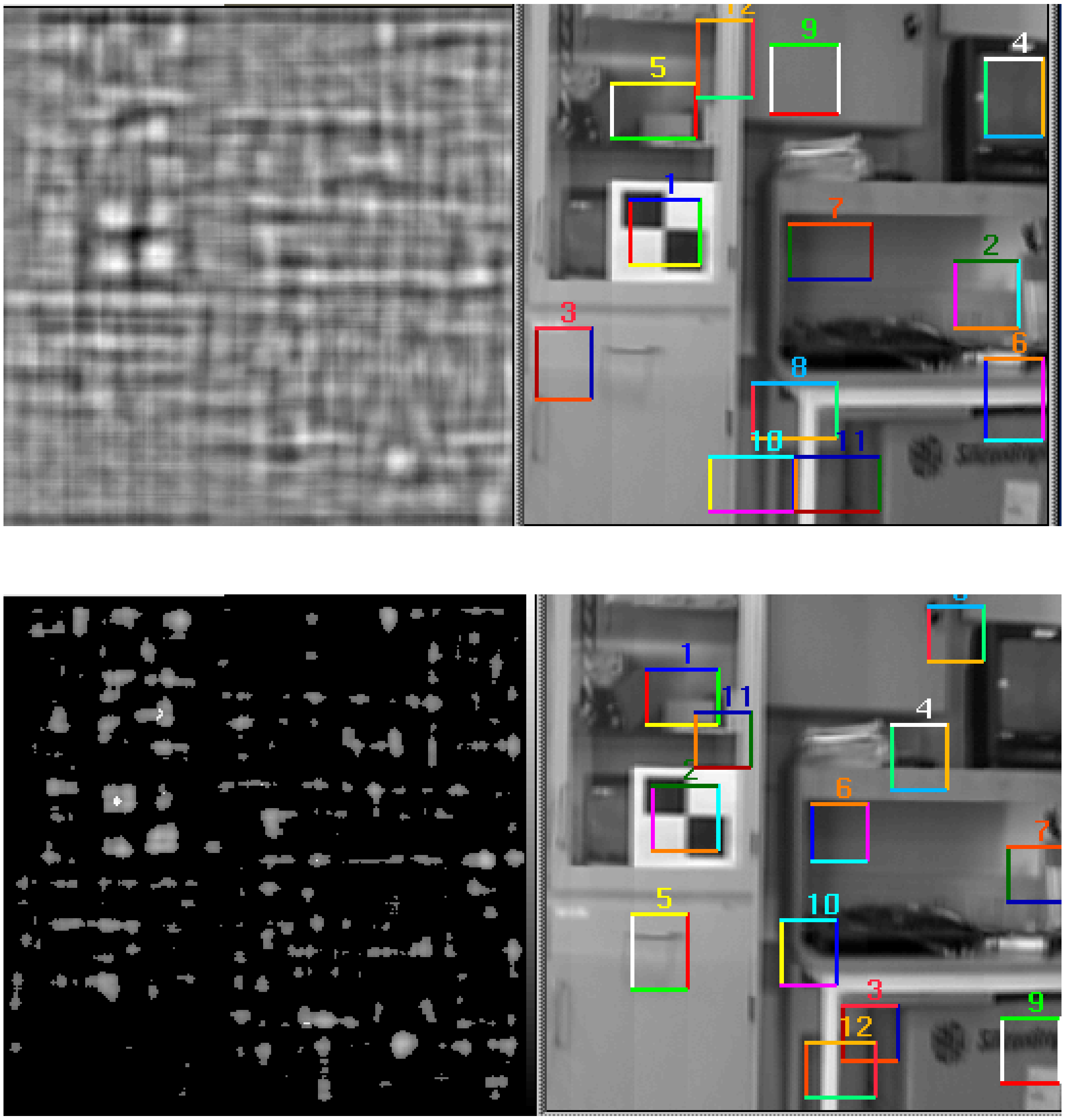


Figure 5.7: The figures show the benefit of using the Difference of Gaussian (DoG) filter. The images at the right-hand side show examples of regions selected after that the proposed attention selection mechanism was run on the feature map (top-row) and on the feature-map filtered by the DoG, i.e. the conspicuity map, (bottom-row). The latter generates more discriminant selected patches and it shows how the ambiguity problem can be reduced by the DoG filter also in case of local-symmetries.

The last feature considered by proposed attentive scheme is the light intensity. When running a DoG filter on intensity images high-contrasted areas are likely to be detected. Despite the intensity feature is not invariant to illumination changes, experiments have shown that if light condition changes moderately, the presence of this feature improves performance of the attention selection mechanism.

The figures 5.8 and 5.9 third-row from the bottom show the result of filtering with DoG feature-maps related to features as corners, vertical edges, intensities and local symmetries. As the figure shows, enhanced areas of interest correspond to regions in the original image which contain the selected feature and with a high contrast.

First, the attention mechanism was tested on each feature at a time. This was made in order to understand how each single feature affects the final selection. The figure 5.8 top-row shows the extracted landmarks by running the Culhane-Tsotsos attention selection mechanism on the four proposed types of image features one at a time.

The experiments showed that if only one feature is considered the results is not satisfactory for our purpose. That is, the result may not be a discriminating texture. For example, when local symmetries were considered as feature alone, the attention selection worked fine only on specialized landmarks, (they were selected). However, the response was not satisfactory for other extracted landmarks, since symmetric regions around horizontal edges were selected. The represented texture patches, only containing horizontal edges, are in fact not acceptable since ambiguous. These landmarks may in fact match the extracted regions as well as the neighbor left and right regions. A similar ambiguity also arises in cases of other image features. For example, near the represented table edges.

Saliency Map

The interesting issues of the above mentioned phase of experimentation was to prove that features need to be integrated. For example, the ambiguity problem described above can be solved by combining different type of features. The idea of utilizing multiple cues simultaneously is very potential. The advantage is that more tasks can be serviced, and hypothesis that depends on co-existence of features may be considered.

In the literature a few different attentive schemes have been proposed to combine different feature types. Features are normally extracted in parallel and then later fused to a unified representation, the *saliency map*. The saliency map is then the input to the attention selection process.

Probably, the most prominent technique to fuse different maps is based on weighted averaging. In the proposed method the different conspicuity maps, Con_i, are fused to a single saliency map, Sal, by summing the individual responses with weight attached.

$$Sal(x,y) = \frac{1}{N}\sum_{i}^{N} \omega_i Con_i(x,y) \text{ , where N is the number of feature maps} \qquad (5.8)$$

The advantage of the process is the simplicity. The disadvantage is that some heuristics need to be employed to set the weights.

The result achieved by including corner, vertical edges, intensity values and circular symmetries as features are showed in figure 5.9 second-row from top which shows a computed saliency map and figure top-row shows extracted visual landmarks.

More in general, the figures in the next two pages represent:

- Figure 5.8 shows through example images the proposed attentive scheme (summarized in figure 5.10). In the represented case the attention selection mechanism was tested on the four proposed types of image features one at a time. The figures top-row consequently represent the extracted texture-patches related to the considered features. It can be noted that the result may not be a discriminant texture.

- Figure 5.9 shows through example images the proposed attentive scheme (summarized in figure 5.10). In the represented case the attention selection mechanism was tested on the four proposed types of image features merged into a saliency map. The figure top-row represent the extracted texture-patches related to all the considered features. It can be noted that the result show the extracted textures are more discriminant than those selected by running the attention scheme on the considered features one at a time.

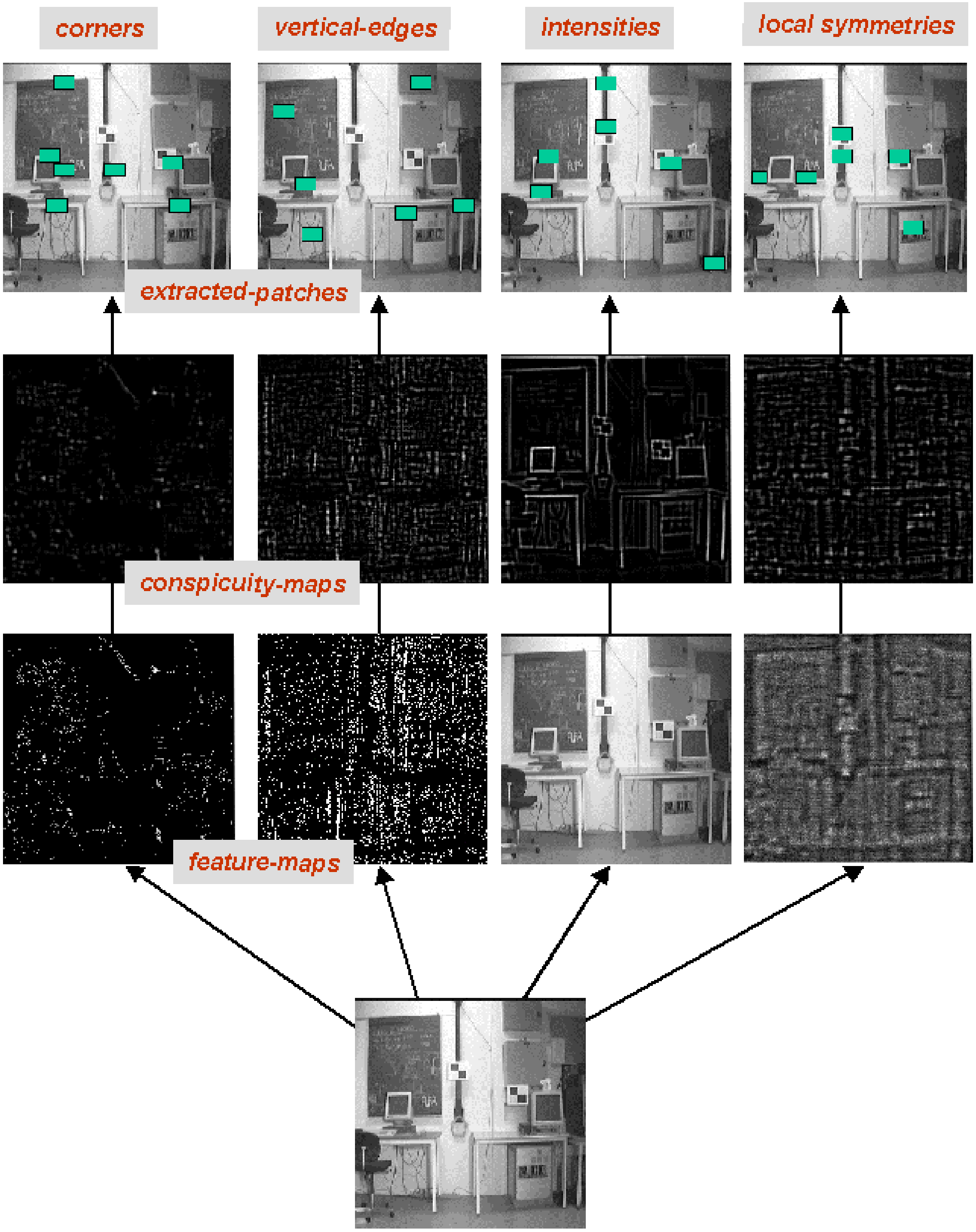


Figure 5.8: The figures show through example images the proposed attentive scheme tested on one feature at a time (see text).

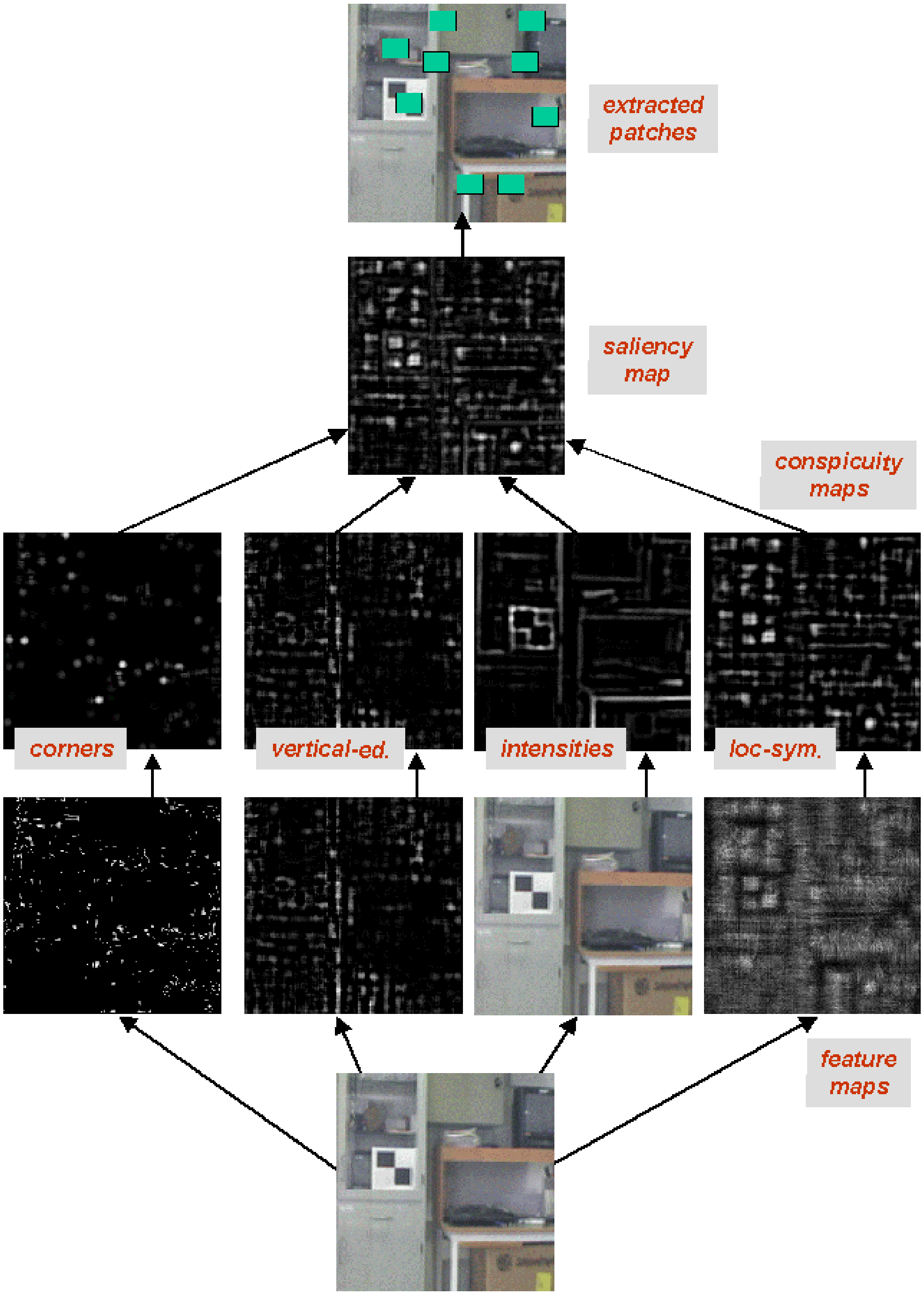


Figure 5.9: The figures show through example images the proposed attentive scheme tested on the four proposed features (see text).

Summary

The proposed attention scheme has consequently been designed as described in the following lines and represented in figure 5.10.

As first step of computation, feature maps are extracted in parallel and processed with Difference of Gaussians filter to emphasize image-regions with a higher number of detected features. Then, the resulting conspicuity-maps are fused into a single saliency map by a weighted average. The saliency map specifies the input to the proposed hierarchical attention selection process, which provides as output a set of selected regions representing visual landmark candidates. Figure 5.11 shows more examples of extracted landmark candidates where features employed in our experiments were corners, intensity values, vertical edges and local symmetries.

The advantage of the proposed method is on using multiple features and on allowing for a fast processing. The proposed attentive process also showed its strength on extracting discriminant feature, unique in the environment. However, few drawbacks were also experienced, for example, there is a difficulty on tuning the attention mechanism for an arbitrary environment.

It has also been experienced that in some cases not all the extracted textures are discriminant. This happened in cases where white walls were predominant in the represented scene. For example, the panoramas showed in figure 5.14 have plenty of white-walls regions. This situation calls for further checks in order to discard unsuitable landmarks. This is one of the reasons why further checks have been introduced as described in next two sections. The other reason for further checks is that they are useful to discard texture patches that, despite being discriminant are unsuitable to become landmarks due to high positional uncertainty or "bad positioning" problems, e.g. they easily become occluded, their texture contains reflections, etc.

In particular, sections 5.2 and 5.3, (titled respectively: Landmark Pose Computation and Model Refinement), will discuss how landmark reliability and discriminant characteristics can furtherly be tested in order to *elect* trustworthy landmark candidates. This is the reason why texture patches extracted by the proposed attentive process are often referred in this thesis as *landmark candidates*. In conclusion, the proposed attentive scheme plays the fundamental role of providing the visual landmark templates for the localization phase.

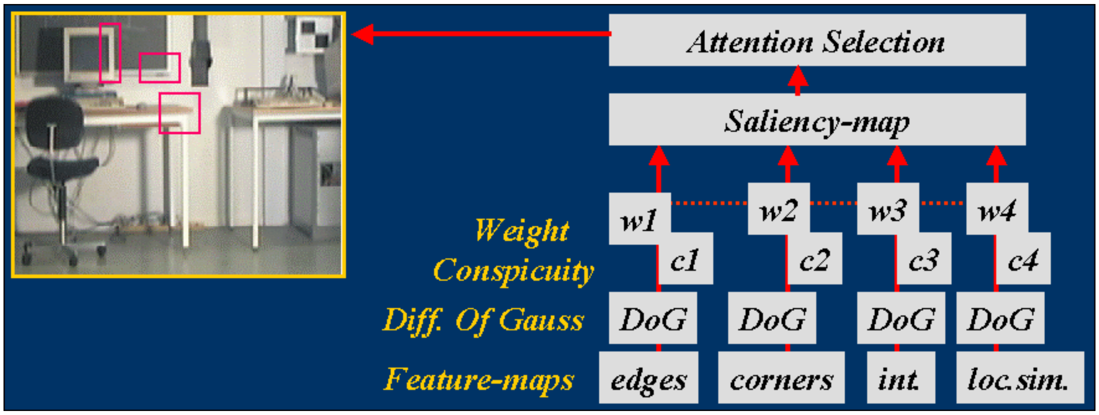


Figure 5.10: The figure shows the proposed procedure for landmark candidate extraction. As first step of computation, *feature maps* are extracted in parallel and processed with Difference of Gaussians filter to emphasize image-regions with a higher number of detected features. Then, the resulting *conspicuity-maps* are fused into a single *saliency map* by a weighted average. The saliency map specifies the input to the proposed *hierarchical attention selection process*, which provides as output a set of selected regions representing *visual landmark candidates*.

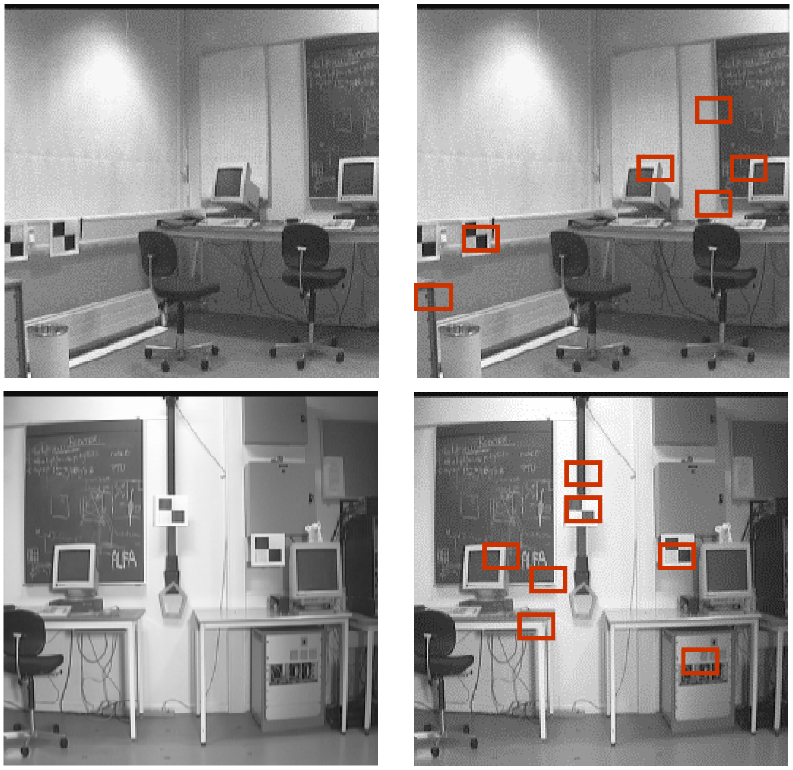


Figure 5.11: Left-hand figures: input images. Right-hand figures: typically visual landmarks extracted, by including corners, vertical edges, intensity values and circular symmetries as searched image-features.

5.1.4 Experimentation

The experimentation concerns with panoramic view synthesis and landmark candidate extraction. The aim was to determine feasibility and performance of the proposed method in a realistic setting. The algorithms were consequently implemented and tested on a real mobile robot system. The hardware used for the experiments was the same as presented in sub-section 4.1.5, (a PC mounted on the mobile platform Robuter by Robosoft, equipped with a monocular camera-head capable of panning over 360 degrees). Figure 5.12 shows the robotic system.

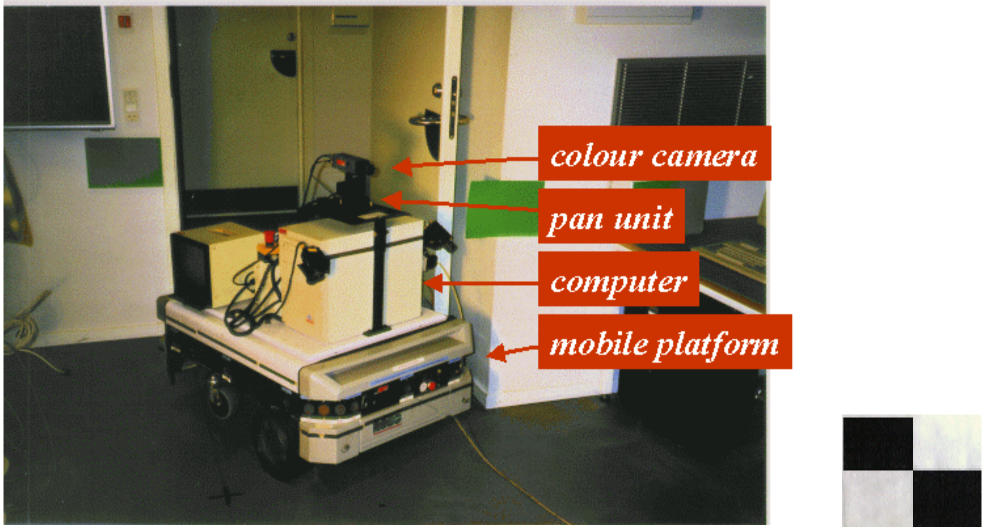


Figure 5.12: The figure left-hand side shows the robotic system used for the experiments. The figure right-hand side shows an image of the specialized landmark introduced in the environment to better test system performance.

The system performance were tested in the two different environments also introduced in chapter 4, (the old and new CVMT laboratory rooms). The workspace floor maps for the two environment rooms are represented in figures 5.13 and 5.14.

How many panoramic-views should be synthesized? and from which position? This topic is discussed in section Learning Navigation Strategy, (5.4). For the present purpose of the experimentation, i.e. test panoramic view synthesis and attentive selection scheme, the position the robot, thus of the camera, was chosen in a way that it could be a typical robot position assumed during navigation of the environment, and at the same time, a position which would allow a large portion of workspace to be visible. The latter in order to provide more data to our experimentation.

Panoramic View Synthesis

During different panoramic acquisition runs of the learning phase, the robot was taken to arbitrary locations. The robot was not moving during the image-acquisition. The need for a stop was due to the proposed acquisition method (subsection 5.1.2) and available hardware. In particular, it was required that the pan unit stopped its rotation every time an image needed to be acquired. This in order to avoid blurred images. The time needed for a 360 degree acquisition and for the synthesis of the panoramic view was around 3.5 minutes. Once the panoramic acquisition ended, the robot can be let to move towards next acquisition position.

In the experiments made in the old laboratory the robot stopped in two locations labeled L and R in figure 5.13. A dense sampling of workspace views were then collected by panning the camera 360° relative to the robot heading. In particular, 72 images were acquired, (i.e. one image every 5 degrees of rotation).

It was decided to consider only the central area of each acquired images in order to discard areas affected by lens distortion. In particular, by considering the available camera lens it was selected an area of 350x350 pixels around the image center. The whole image resolution was 512x512 pixels. Successfully we experimented in the old-laboratory with building panoramic views using cylindrical projections. Figure 5.13 middle and bottom rows show synthesized panoramic views related to position L and R respectively. The positions L and R are represented in the floor-map of figure top-row.

The overlapping area between two adjacent images was 85.7% and each pixel was averaged over 7 different images. Because of the dense sampling and the averaging process, it is not possible to distinguish among images inside the panoramas. The only visible effect is that objects may appear slightly unfocused and this has the effect that the image become a little blurred. This effect had no apparent consequence in the next computation steps. However, in later experiments a median filter was used to merge correspondent pixels since it better preserves edges.

In the experiments made in the new laboratory the robot stopped in five locations labeled A_l, A_r, B_l, B_r, C in figure 5.14 top-row. The sampling rate was chosen as for the previously described panoramic acquisition since the successful result. However, it was decided to reduce the central area of each acquired images from 350 x 350 to 256 x 256 to further decrease the effect of lens distortion, (the whole image resolution was 512 x 512 pixels). The panorama final resolution was then 256 x 4101 pixels. Once again, panoramic views were successfully synthesized using cylindrical projections. Figure 5.14 shows examples of a synthesized panoramic views, and workspace floor-maps including the acquisition positions.

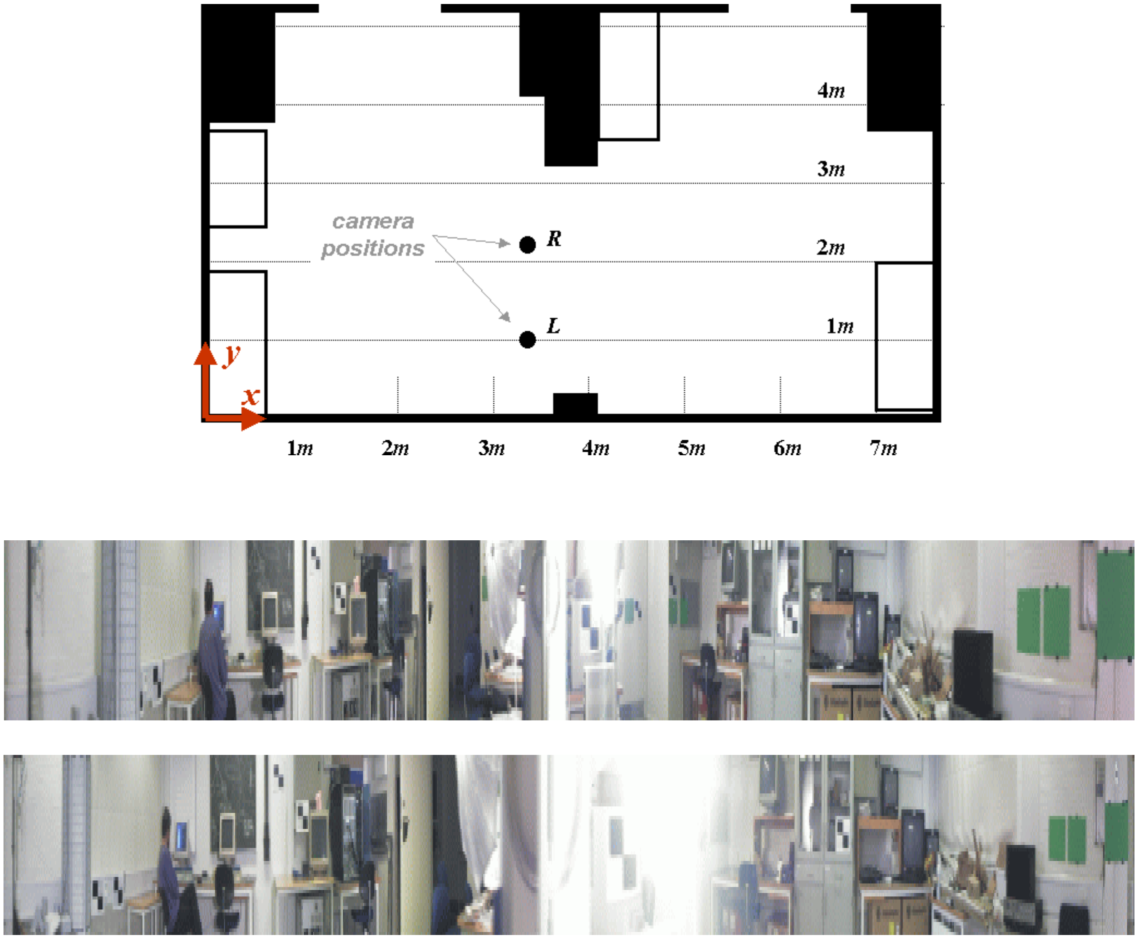


Figure 5.13: The figure top-row shows a floor-map of the old laboratory. The dark dots labeled L and R represent two locations chosen for the acquisition of panoramic views. The figure bottom-row shows panoramic views related to position L and R.

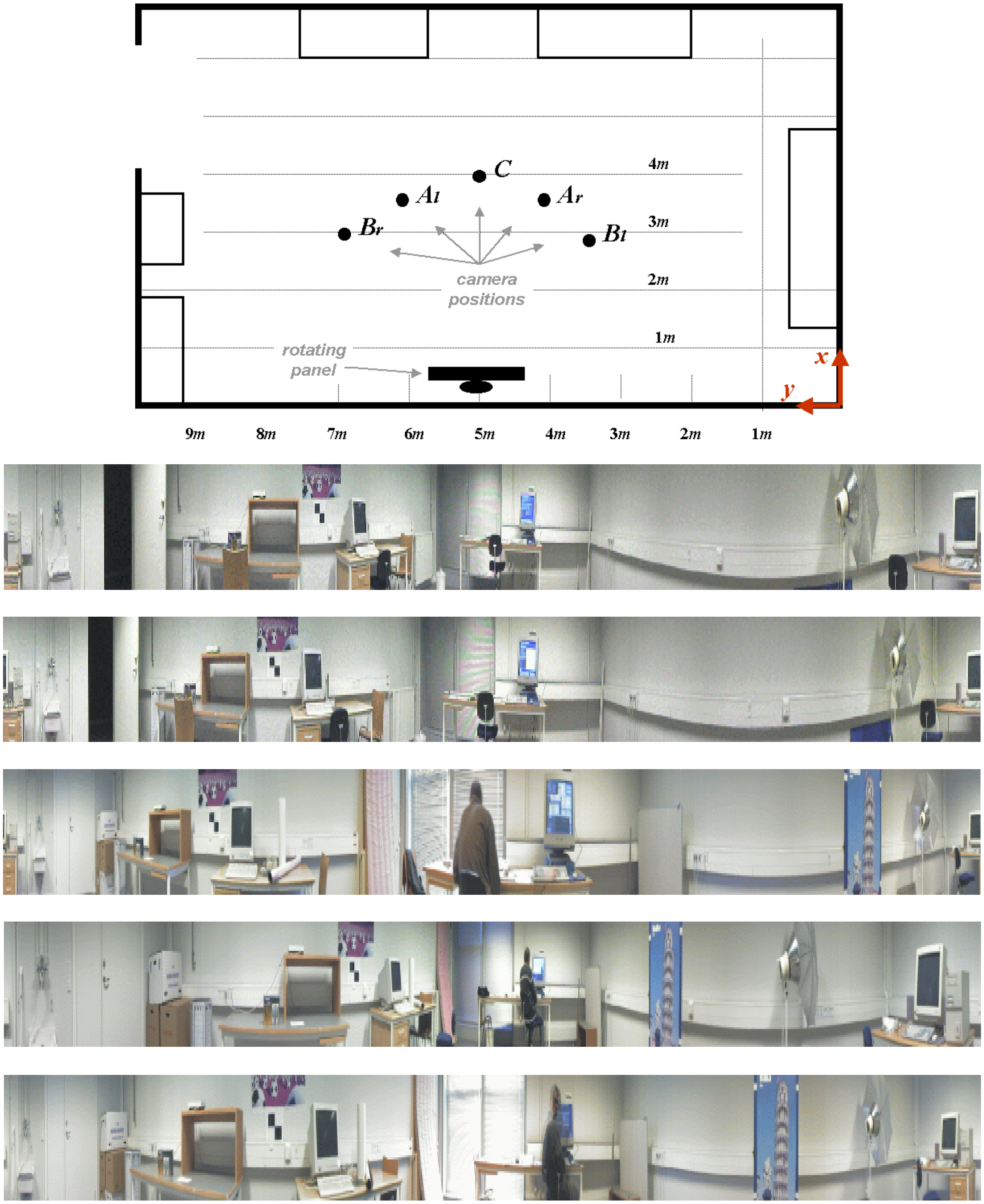


Figure 5.14: The figure top-row shows a floor-map of the new laboratory. The figure shows the five chosen locations chosen for the acquisition of panoramic views, they are labeled A_l, A_r, B_l, B_r, C. The other figures show the panoramic views synthesized from the above mentioned five locations respectively (in different periods of time).

Landmark Candidate Extraction

Once a panoramic-view has been synthesized, the proposed attentive scheme for landmark views extraction can run on it. The attentive scheme was only run on specific sub-panoramic regions typically sized 512 columns each, and related to areas of the workspace considered most suitable for landmarks. This followed the argumentation given in section 5.4 about the *principal viewing axis*, (see also Yuen and MacDonald [150]).

In order to provide the workspace with more interior characteristics of typical indoor environments, posters were introduced in the new laboratory. The previously introduced specialized landmark, i.e. textured as in figure 5.12 right-hand side, was also introduced both on walls and on the poster-pictures in order to test discrepancy in performance between ideal and typically selected landmarks.

During the experimentation the robot was taken to the seven locations showed in the floor maps of figures 5.13 and 5.14 top-rows. The attentive scheme used in our experiments was implemented having corners, vertical edges, intensity values and local symmetries, as the searched image-features. Figure 5.10 described the attentive computational scheme, and figures 5.15 and 5.16 show examples of visual landmarks extracted on sub-panoramas related to the old and the new laboratory respectively.

A variable number of texture patches was required to be extracted by the system depending on the workspace (natural) texture. In case of the old laboratory, the extracted landmarks were in average between 12 and 29. Only 15 texture-patches were instead extracted from the panoramas related to the new laboratory. This happened because of the lack of textures in the room and the smaller size of sub-panoramic images, (256 rows instead of 350).

The experiments were also performed to test extraction of landmarks having different orientations. The results showed that landmark selection depends on the landmark orientation but also on the distance to a landmark. In particular, the distance and orientation to a landmark affect shape and size of the projected region in the camera image-plane. When projection area becomes significantly smaller than the visual template "ideal-size", (the "ideal size" was estimated between 28x28 to 36x36 pixels), the landmark is not selected any longer. Figure 5.17 top-row shows the workspace setup through a panoramic view. The environment was left quite empty of furniture but filled in with plenty of ideal landmarks patterns.

It was observed that the "ideal" landmark patterns were not extracted when their orientation to the camera, or their distance, generated in the camera-image an insufficiently large "projection-region". Namely, smaller than 30% of the "ideal" template size. Figure 5.17 (three rows from the bottom) shows typical results. The ideal landmark is always selected when its dimension in the image-plane approximately occupy over 70% of the template ideal size. If the re-projected region is larger then the ideal size a part of ideal landmark is always selected.

Other runs of experiments were made to confirm the proposed landmark ideal size. In particular, template size were increased up to twice the ideal template area. Figure 5.18 bottom-row shows extracted landmarks in comparison to ideal size (top-row). Extracted landmarks were for most of the cases the same as extracted with the ideal size setup, which confirmed reliability of the ideal size. Note that, as previously discussed in section 5.1.1, despite bigger templates make more robust their recognition, they are more sensitive to changes in viewpoint of observation (new visible aspects may arise and occlusion may be generated). Consequently, a bigger template size is only advised to be used when potential occluders, (e.g. the tables and the computer in figure 5.18), are distant from extracted landmarks. It was also experienced the importance of tuning the selection mechanism on the type of texture expected for a certain environment, (as well as typical distances to landmarks).

Note that not all the extracted texture-patches represented discriminant textures. This can be observed in figure 5.16. Nevertheless, the discriminant characteristics of the selected texture patches will furtherly be tested later in the learning process, for example, by comparing acquired texture patches from different view points, etc., but this will be discussed in next two sections.

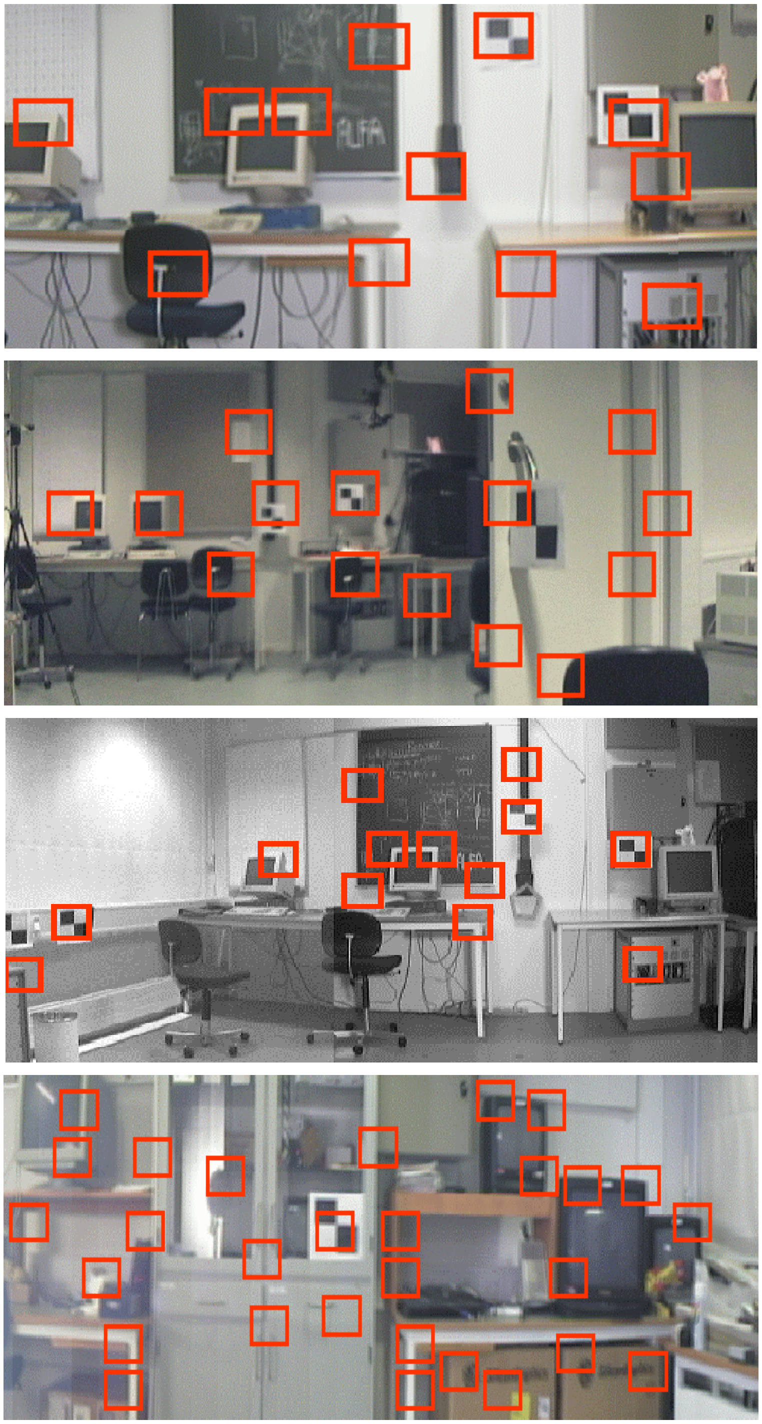


Figure 5.15: The figures show examples of visual landmarks extracted on sub-panoramas related to the old laboratory.

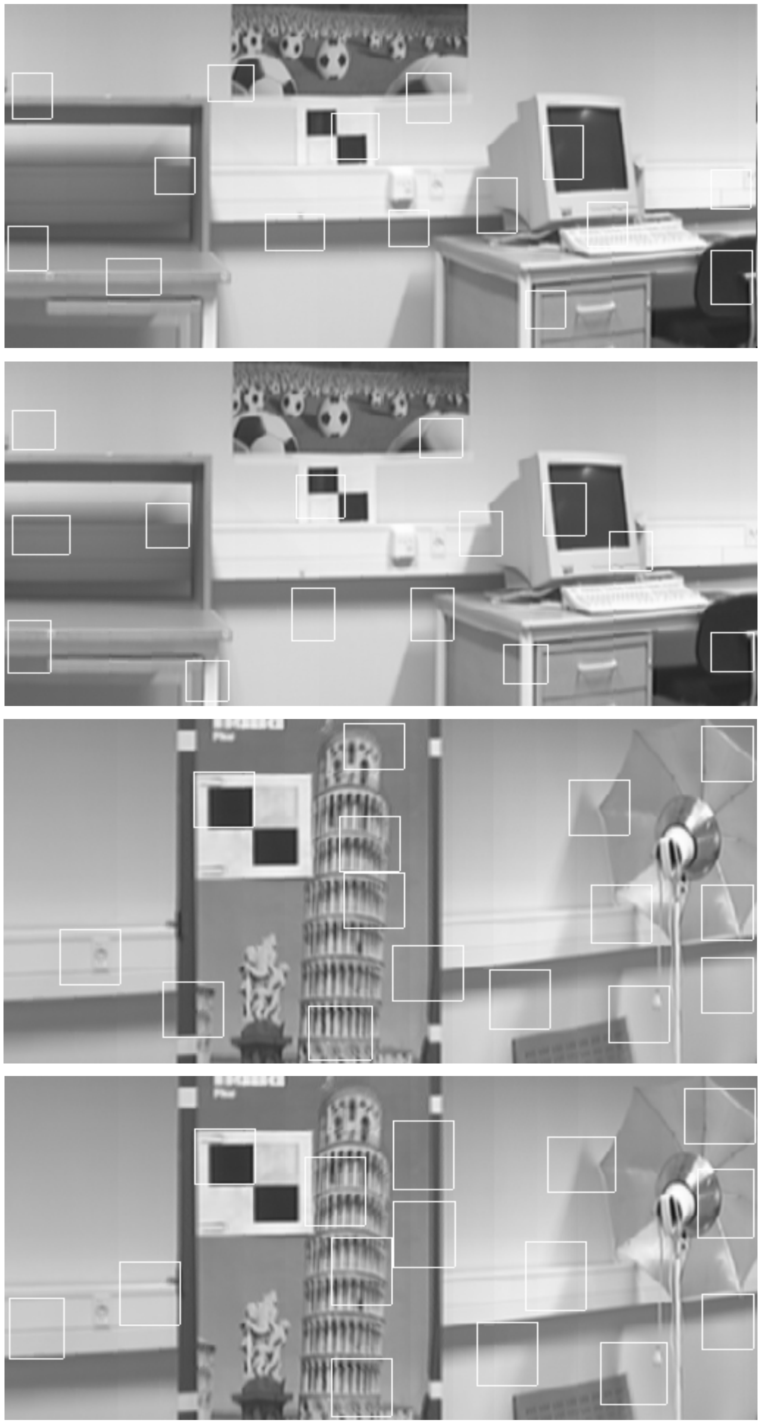


Figure 5.16: The figures show examples of visual landmarks extracted on sub-panoramas related to the new laboratory.

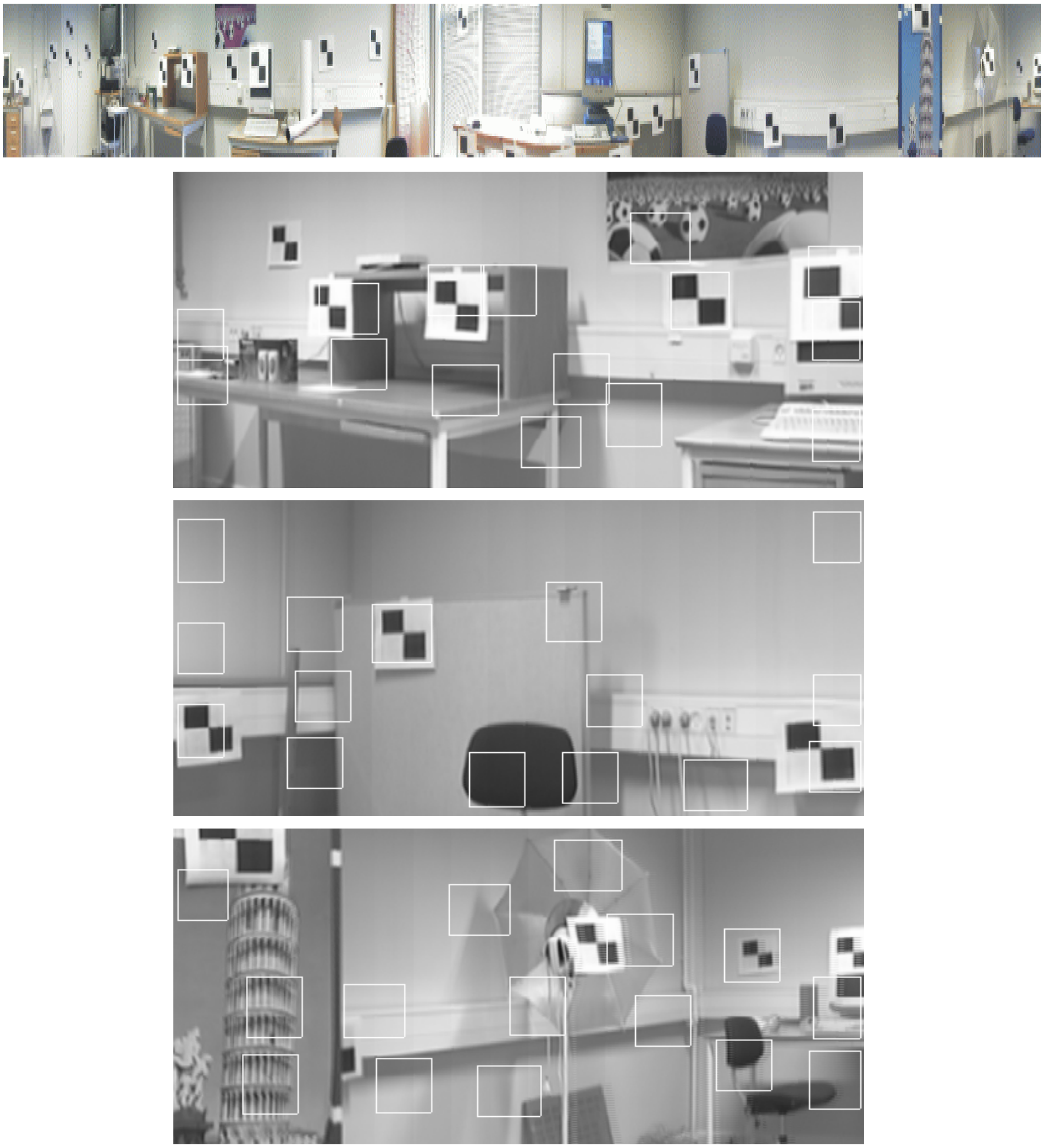


Figure 5.17: The figures show results of experiments aimed to test extraction of landmarks having different orientations. To this purpose the workspace was filled in with plenty of "ideal" texture patterns. The figure top-row shows the experimental setup through a panoramic view. The figures in the three rows from the bottom shows typical extracted landmarks. The ideal landmark is always selected when its dimension in the image-plane approximately occupy over 70% of the template ideal size. If the projected region is larger then the ideal size a part of "ideal" landmark is always selected.

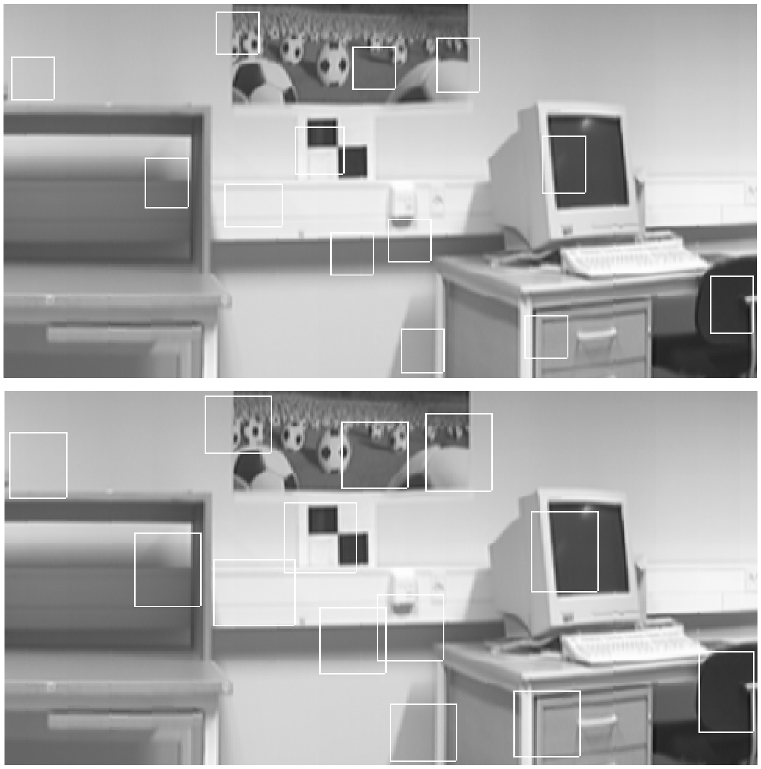


Figure 5.18: The figures show results of experiments aimed to confirm the proposed ideal size by testing extraction of landmarks having different texture-template size. In particular the figure bottom-row shows the case when the template size is increased up to twice the ideal template area. Extracted landmarks were for most of the cases the same as extracted with the ideal size setup, which confirmed reliability of the ideal size.

5.2 Landmark Pose Computation

The landmark positional knowledge is relevant to many applications which need metric navigation and accurate pose estimation. In case of the proposed system, knowledge of landmark pose is necessary to the self-localization process in order to:

1. project previously learned landmarks from the current camera viewpoint, so allowing for their recognition in the camera image-plane, (see section 4.1);
2. estimate the absolute robot pose in the workspace, based on the triangulation method, (see section 4.2).

Due to the requirement of accurate metric navigation, landmark pose in the workspace needs to precisely be estimated. Special attention must consequently be paid to the error associated with estimated landmark poses and to how this error will affect robot localization performance.

The knowledge of landmark pose will allow the system to continuously re-compute the absolute robot pose during navigation, making the system tolerant to positional errors arising from previous pose estimations, i.e. errors do not accumulate, [93].

5.2.1 Core Idea and Argumentation

Full autonomous navigation implicates the capability of navigating unknown environments. Consequently, landmark positional information can not be proposed as an a priori knowledge (for example, a map of the environment), and the landmark position needs to be autonomously learned by the system. In our case the system needs to learn position of environment characteristics represented by the texture patches which have been extracted by the method presented in section 5.1.

The only information the system possesses about the extracted patches is their location in camera image plane. Nevertheless, the system also possesses knowledge of camera parameters and the robot poses at the time textures have been captured.

The 3D position of environment characteristics represented in the texture patches can in principle be inferred by vision-based stereo reconstruction. A task that would instead be impossible to perform using some alternative sensor modalities like the sonar. As for the laser modality, this could represent a possible solution in order to estimate 3D position of identified image-regions, (supposing an opportune calibration between laser and camera system). However, the laser option is not investigated in this thesis, since it has been decided to only rely on the more affordable visual stereo reconstruction and so investigate how well such a solution may perform.

The stereo reconstruction technique is known to be very challenging due to the well known correspondence problem. The uncertainty related to camera external parameters plays an important role (as well as the texture characteristics of corresponding landmarks). As for the camera internal parameters, it is proposed they should be estimated previously and precisely, by an off-line procedure. A careful analysis of proposed reconstruction techniques should consequently be addressed to understand if an accurate and reliable landmark position estimation, based on a visual stereo reconstruction, can be achieved by the proposed system.

The extracted texture patches may represent any object in the environment, but also more than one object or just a part of one of more objects. Consequently, the 3D surface representing a landmark may be of any 3D shape. Modeling arbitrary 3D shapes by stereo reconstruction can easily lead to errors which would make the computed output imprecise and unreliable. It is then suggested to simplify the reconstruction process by setting a range of possible shapes that a landmark can have. For this thesis work, it is proposed to only consider *planar landmarks*, i.e. sub-images representing world planar surfaces. Note that planar landmarks represent a popular choice in this field. The system is then required to be able to detect landmark planarity.

Provided that planarity can be detected we get these advantages:

1. The 3D reconstruction of planar landmarks based on stereo vision may become a simple and very reliable process. This aspect is very useful to estimate landmark pose during the learning phase.
2. Landmarks may more easily be identified by different viewpoints than any-shape landmarks. In fact, the landmark surface would not create occlusions to parts of the same landmark when the viewpoint changes, (still, landmarks can be occluded by other objects). This aspect is useful in order to find landmark correspondences during the learning phase.
3. Landmarks are suitable to re-projection, (especially when the reprojection is based on representative views). In fact, landmark views do not posses hidden parts which may not be recreated when observed from new viewpoints. This aspect is very useful in order to recognize previously acquired landmarks during the localization phase.
4. A typical indoor environment is plenty of planar surfaces which could be used as landmarks.

Since a landmark will represent a portion of a planar surface, the positional information to be learned by the system can be represented by the position of landmark center and orientation of landmark planar surface.

In this thesis, only landmarks representing vertical surfaces are considered. This choice is in order to reduce the inherent distortion in acquired textures arising from inconvenient relative orientations between landmark and camera. In fact, due to the facts that the robot moves on a planar surface and that the only allowed camera motion is the panning. The camera image plane will consequently always be vertical, which leads to the fact that landmark surfaces which are not vertical, will always contain some degree of distortion. Provided that the orientation of the surface can be detected, this will only be accepted around the vertical axis.

The positional information represented by the *landmark pose* can consequently be represented by: 3D position of landmark center (x, y, z) and orientation around the vertical axis of landmark surface, (θ_z). Figure 5.19 shows a representation of robot workspace including global coordinate axis and an example of robot and landmark pose.

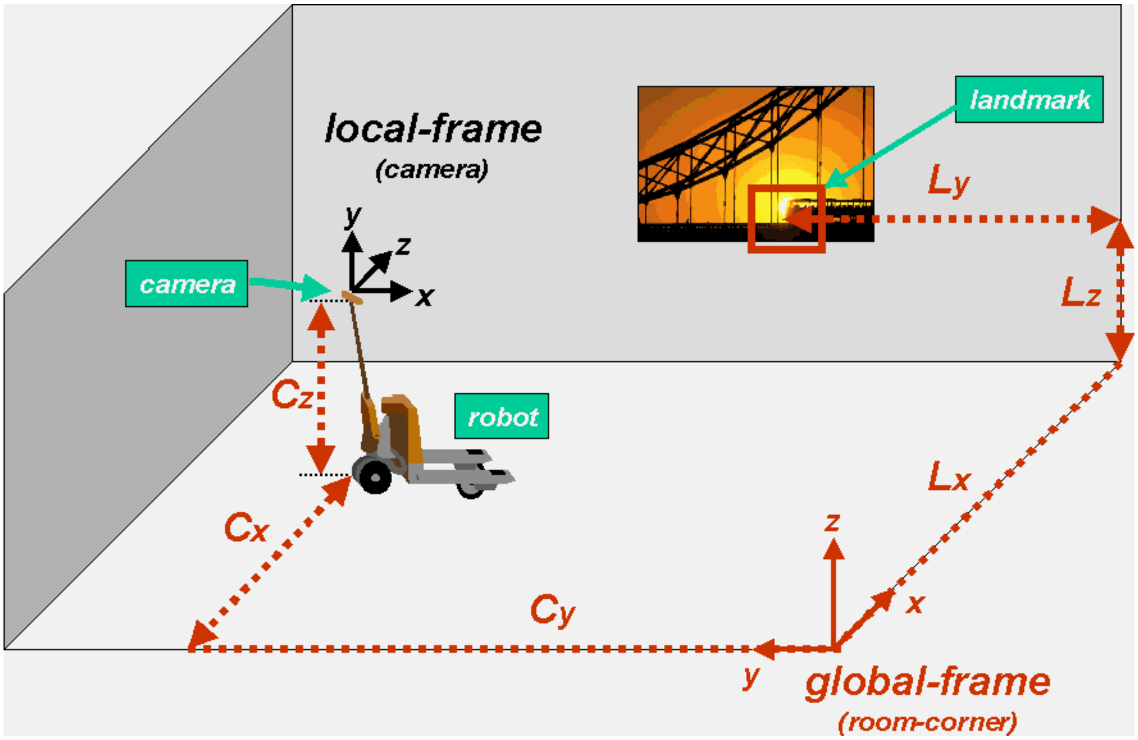


Figure 5.19: The figure shows a representation of the robot workspace including global and local coordinate axis, an example of robot, and a landmark represented by a portion of a poster hanging on a wall. The represented robot, (a fork-lift like those used in warehouses for transporting and positioning of heavy material), is imagined to be equipped with a camera on its top sitting on a rotating support. This kind of robot could thus be suitable for adopting a localization system based on the method proposed in this thesis.

The process of learning landmark positional information is consequently focused on:

1. estimation of landmark central position;
2. recognition of landmarks representing planar surfaces;
3. estimation of landmarks surface orientations.

But, what do we need to make the process of learning landmark pose, accurate and reliable?

The answer to this question is the method presented in the following sections. In particular, sub-sections Correspondence Test (5.2.2) describes the proposed way of finding correspondences between extracted landmarks into different views, as well as a way of testing information reliability; and Pose Extraction (5.2.3), describes the method proposed to estimate landmark pose. Last sub-section, Experimentation (5.2.4) will then present obtained results.

5.2.2 Correspondences Test

Extracted landmark views are required to be representatives, that is, they should clearly represent all texture details of represented objects and they should be taken close enough to allow for a fidelity re-projection from new viewpoints. This topic has previously been discussed (sub-section 5.1.2).

Naturally, it would not be practical to collect perfect representative views of landmarks so that a compromise solution has been set. This is represented by a view point acquisition which is *"close to frontal"*, (i.e. frontal within a certain degree), and which does not involve major occlusions. The above favorable condition can be identified by observing the same landmark from different view points. In particular, if different observations of the same workspace region match each other, the represented texture is considered representative to that region.

Two questions consequently arise: (1) What method should be adopted for matching different views related to the same workspace region? (2) How representative can a matched view be considered?

The answer to the first question is addressed in the following "Normalized Cross-Correlation" and "Cylindrical Epipolar Geometry", while the second issue is discussed in sub-section, Learning Navigation Strategy (5.4). The latter is discussed after landmark "pose extraction" because the knowledge of landmark pose may be used to estimate how representative is a matched view.

Normalized Cross Correlation

There are many algorithms proposed in the literature concerning image matching. In general, the comparison between two images can be feature or correlation based. Analogously to what was chosen for locating landmarks in the image-plane for self-localization, the correlation based technique is preferred for its characteristic of reliable texture recognition, (see sub-section 4.1.3). The formula related to normalized cross correlation technique is shown below:

$$NCC(x,y) = \frac{\sum_{v=0}^{Vmax} \sum_{u=0}^{Umax} R(u,v) \cdot S(x+u, y+v)}{\sqrt{\sum_{v=0}^{Vmax} \sum_{u=0}^{Umax} R^2(u,v) \cdot \sum_{v=0}^{Vmax} \sum_{u=0}^{Umax} S^2(x+u, y+v)}} \tag{5.9}$$

where $R(u,v)$ is the landmark visual template, $S(x,y)$ is the search window, $R^2(u.v)$ is the energy of the landmark visual template, $S^2(x,y)$ is the energy of the search windows, and $Vmax$ and $Umax$ represent maximum template size.

The performance of normalized cross correlation has been tested through an extensive number of runs, (see section Experimentation). The conclusion was that the normalized cross-correlation can be trusted on the 93% of the cases.

Cylindrical Epipolar Geometry

Epipolar geometries are commonly used in the computer vision community for recovery of depth by stereo reconstruction, (Trucco and Verri [142]). The basis for the epipolar geometry is the plane containing two camera optical centers and the workspace point of interest. The epipolar relation then allows for associating the workspace point projection in the first image to a corresponding line in the second image. The existence of this invariant reduces the search for corresponding points from 2D to a 1D problem.

In the case of this work correspondences must occur along *cylindrical epipolar lines*. The cylindrical epipolar line represents a geometric constraint for cylindrical projections that determines the possible positions of a point given its position in some other cylinder. This constraint plays the same role that the epipolar geometry plays for planar projections, [103].

In particular, every point in the considered cylinder corresponds to a ray in space as given by the cylindrical epipolar relation represented in equation 5.10. When one of the rays is observed from another cylinder, its path projects to a curve which appears to begin at the point corresponding to the origin of the considered cylinder, and it is constrained to pass through the points image on the other cylinder.

Cylindrical projection, however, do not preserve lines. In general, lines map to quadratic parametric curves on the surface of a cylinder. The paths of these curves are uniquely determined sinusoids. This cylindrical epipolar geometry has been computed as in [103] using the following equation:

$$W_v^{right}(\gamma)' = \frac{M_x \cos(\tau-\gamma) + M_y \sin(\tau-\gamma)}{M_z} + C_v \tag{5.10}$$

where

$$\begin{bmatrix} M_x \\ M_y \\ M_z \end{bmatrix} = \left(\begin{bmatrix} C_x^{left} \\ C_y^{left} \\ C_z^{left} \end{bmatrix} - \begin{bmatrix} C_x^{right} \\ C_y^{right} \\ C_z^{right} \end{bmatrix} \right) \times \begin{bmatrix} \cos(\tau - W_\gamma^{left}) \\ \sin(\tau - W_\gamma^{left}) \\ C_v - W_v^{left} \end{bmatrix} \tag{5.11}$$

$(W_\gamma^{left}, W_v^{left})$ represents the angle γ and the ordinate v of landmark-center location in the left-panorama image-plane, $(W_\gamma^{right\prime}, W_v^{right\prime})$ represents the correspondent ordinate v for a certain angle γ, in the right panorama, τ is the rotation offset which aligns the angular orientation of the cylinders to a common frame, and C_v is the ordinate v of the scan-line where the center of the projection would project onto the scene, (i.e. the ordinate of the line of zero elevation). $(C_x^{left}, C_y^{left}, C_z^{left})$ and $(C_x^{right}, C_y^{right}, C_z^{right})$ represent respectively the left and the right camera positions.

The above equation gives a concise expression for the curve, $W_v^{right}(\gamma)'$, formed by the projection of a ray across the surface of a cylinder, (labeled "right"), where the ray is specified by its positions on some other cylinder, (labeled "left"). The cylindrical epipolar curve can be completely specified with no more information that it was needed for the planar case. The geometry of cylindrical epipolar lines is illustrated by the example in figure 5.20.

Figure 5.21 bottom-row shows the cylindrical epipolar line, (the dot-line), superimposed to its related panoramic view. The cylindrical epipolar line is related to the "ideal" landmark.

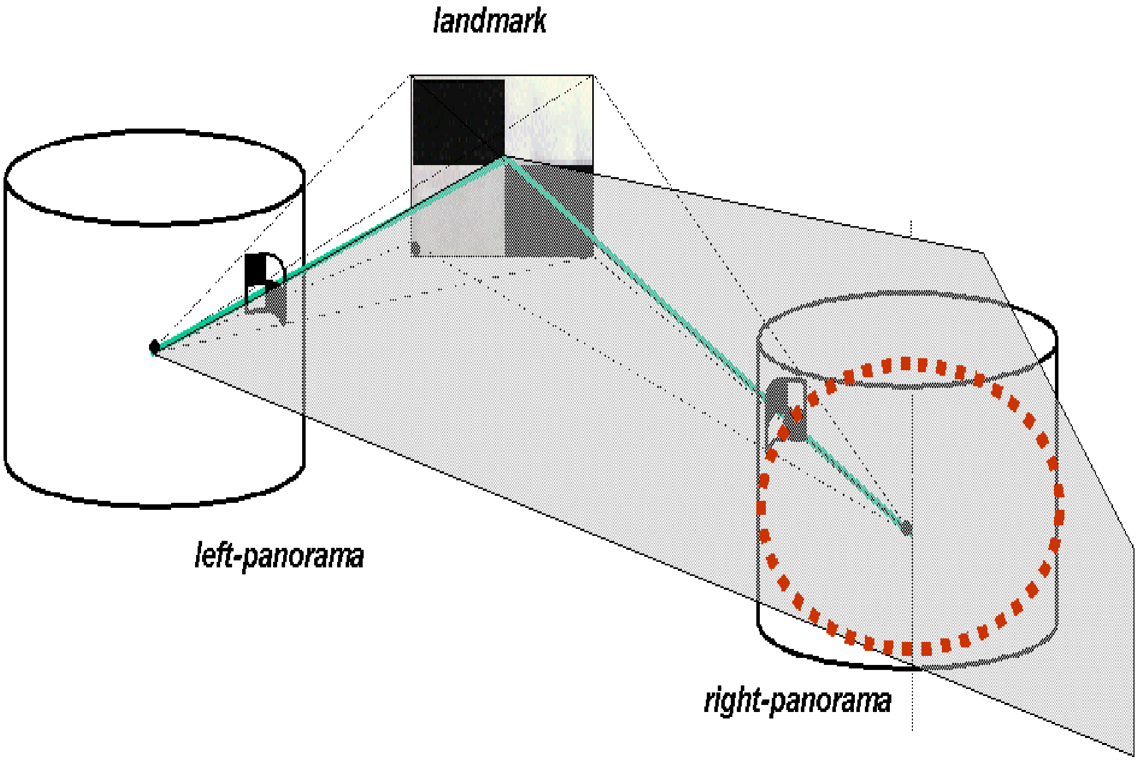


Figure 5.20: The geometry of cylindrical epipolar relation illustrated through an example. The epipolar plane contains the lines connecting the cylinder center of projections to the landmark center.

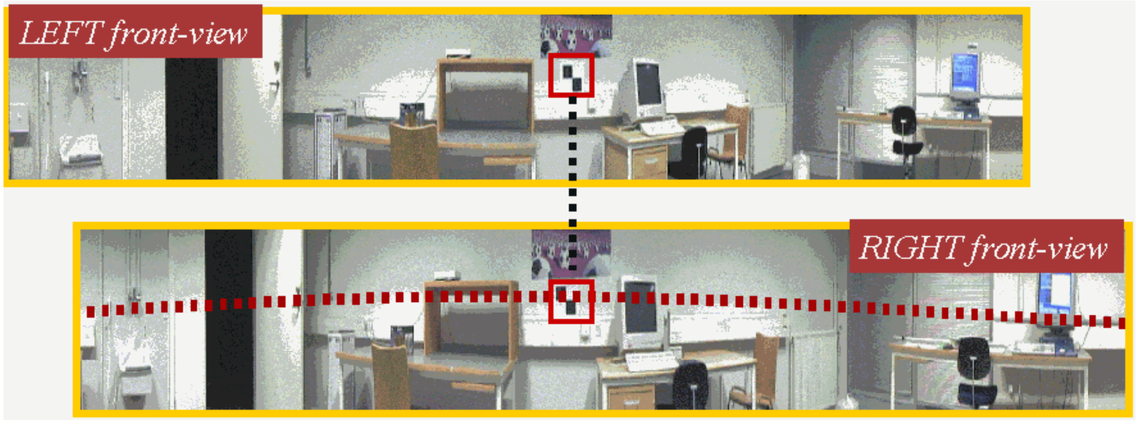


Figure 5.21: The figure shows two computed panoramic views of the robot workspace, (new laboratory). The figure bottom-row shows the cylindrical epipolar line, (the dot-line), superimposed to its related panoramic view. The cylindrical epipolar line is related to the "ideal" landmark represented in figure top-row.

Searching along epipolar lines makes the stereo correspondence computation faster since the search only takes place in a prescribed area along the line. Figure 5.22 bottom-row shows an example of search area, (the thick line), for the epipolar line represented in figure 5.22 top-row. The surface-width of a search area depends on the expected error in landmark location in the camera image-plane, (and other things).

Let the visual template size be R_w x R_h and the size of the search-window, S_w x S_h. The size of the search-window is then set to $S_w = 2R_w$ and $S_h = 1.5R_h$. As for the *incremental step*, that is, the distance between two consecutive search windows along the epipolar line, this is set to R_w. The proposed incremental step, R_w, has been chosen in order to set the precise pixel-distance which would allow two consecutive search windows to identify twice any landmark template lying in-between the two search windows.

Such an incremental step might be seen as a waste of computational time because each valid matched area is probably going to be checked twice (into two consecutive search windows). However, the proposed technique has shown to be very useful in order to reject unsuitable landmark templates as well as confirm valid matches. In particular, a match which is found twice in the same position in two consecutive search windows (along an epipolar line), is very likely a valid match. This was experienced for most of the successful matches. On the contrary, any match which is not confirmed by a consecutive match despite it could be, is discarded since in this case it is possible that the result will be an ambiguous texture patch.

Other than allowing for a faster computation, searching along epipolar lines makes the correspondence problem much more reliable since it prevents from false matches which could have been found outside the prescribed searching area. In addition, based on the geometry of the stereo situation, it is sufficient to only search along a portion of epipolar line which should corresponds to the regions of workspace which are potentially most suitable to landmarks, (see section 5.4).

Experiments demonstrated that searching along epipolar lines allowed for identification and

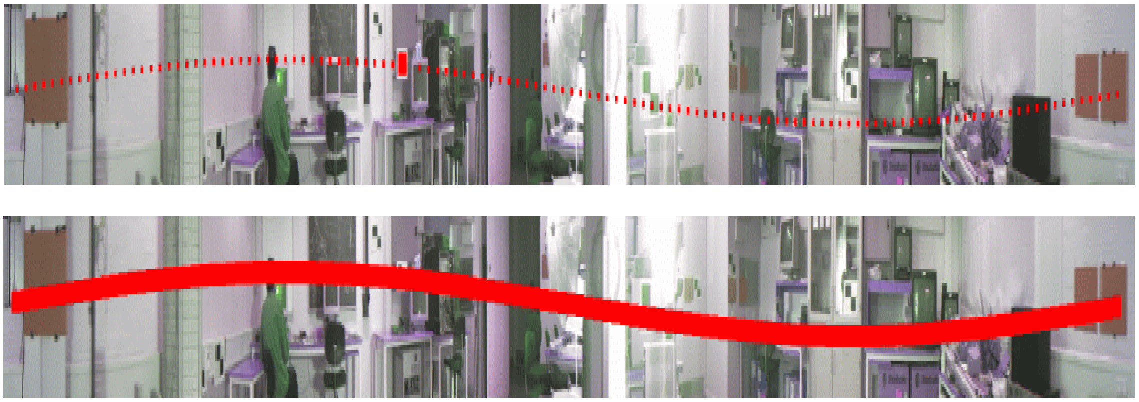


Figure 5.22: The figure shows two computed panoramic views of the robot workspace, (old laboratory). The figure bottom-row shows an example of search area, (the thick line), for the epipolar line represented in figure top-row.

rejection of ambiguous and poorly textured landmarks. This was achieved by using normalized cross-correlation and checking for multiple matches along the epipolar-lines. Figure 5.32 represents an example where unsuitable landmarks were successfully discarded.

5.2.3 Pose Extraction

In order to compute landmark pose, extracted visual landmarks need to be identified into at least two panoramic views. Hence, position of landmark center can be estimated by a stereo reconstruction, and landmark surface orientation can be inferred based on multiple positional information related to a landmark.

Landmark 3D position

The techniques proposed for the correspondences test presented in the previous sub-section, (normalized cross-correlation and search along epipolar lines), can well be exploited to compute landmark 3D position. In particular, texture patches extracted on a panoramic-view by the attentive process presented in section 5.1, can be matched to a new synthesized panoramic-view (from a different view point), by searching along epipolar lines, and by using normalized cross-correlation to detect correspondent patches.

To simplify procedure explanation let us hypothesize that extraction of texture patches takes place in the "left" panoramic view, while a match is searched in the "right" panoramas. For each successful match, the position of landmark center in the workspace can be estimated by stereo reconstruction.

The geometry of the 3D reconstruction is computed using the equations 5.12 for the horizontal plane and the equations 5.14 for the vertical plane. The reconstruction geometry on both horizontal and vertical planes is represented in 5.23 middle and bottom row. The local coordinate system has its origin lying in the mid-point of the baseline b, (illustrated in figure 5.23 top-row). The baseline is defined as the distance between left and right camera optical centers. The equations in 5.12 compute the landmark x and y coordinates in the local coordinate system. In particular:

$$l_x = \frac{-(b/2)\cdot\sin(\theta_1-\theta_2)}{\sin(\theta_1+\theta_2)} \qquad l_y = \frac{b\cdot\sin(\theta_1)\cdot\sin(\theta_2)}{\sin(\theta_1+\theta_2)} \tag{5.12}$$

where $\vec{l}=(l_x,l_y,l_z)$ is the 3D position of landmark center in the local coordinate system, and the angles θ_1 and θ_2, are the angles between the baseline and the line connecting camera optical-center and landmark-center as shown in figure 5.23 bottom-row. These angles and the baseline b can be obtained by the following equations.

$$\{\begin{matrix}\theta_1 = c_\theta^{left}-\phi_1 \\ \phi_1 = \arctan(w_u^{left}/f)\end{matrix} \quad ; \quad \{\begin{matrix}\theta_2 = \pi - c_\theta^{right}+\phi_2 \\ \phi_2 = \arctan(w_u^{right}/f)\end{matrix} \quad ; \quad b = c_x^{left}-c_x^{right}; \tag{5.13}$$

where $\vec{c^{left}}=(c_x^{left},c_y^{left},c_z^{left},c_\theta^{left})$ and $\vec{c^{right}}=(c_x^{right},c_y^{right},c_z^{right},c_\theta^{right})$ are respectively the left and the right camera poses in the local coordinate system. Note that in the local coordinate system $c_y^{left}=c_z^{left}=0$ as well as $c_y^{right}=c_z^{right}=0$. The locations of a landmark into the left and right camera image-planes are respectively described by $\vec{w^{left}}=(w_u^{left},w_v^{left})$ and $\vec{w^{right}}=(w_u^{right},w_v^{right})$, where $w_u^{left}=w_\gamma^{left}(\frac{"num-pixel-in-panorama-horizontal"}{2\pi})$, (analogously for w_u^{right}). The angle between camera optical axis, and the lines connecting camera

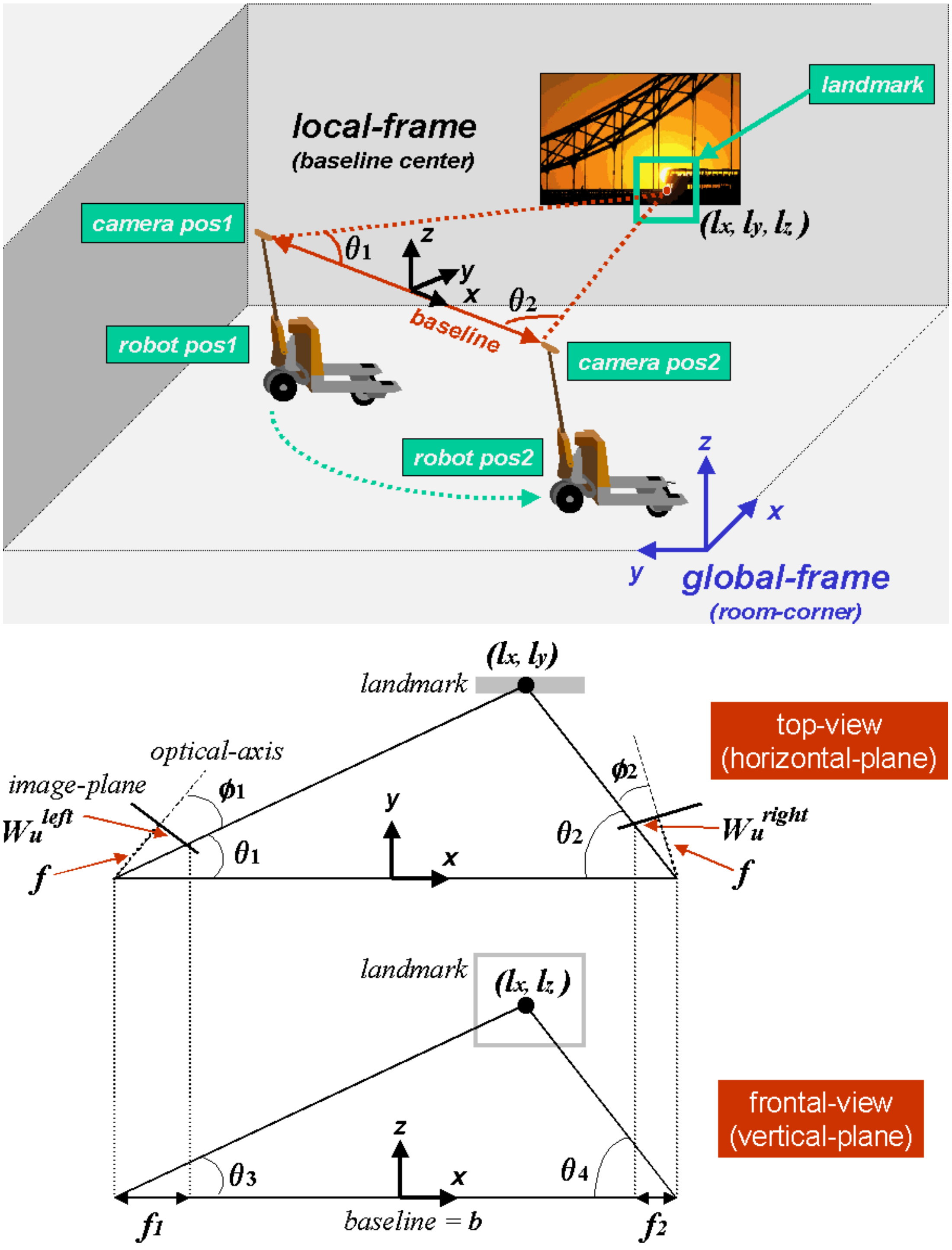


Figure 5.23: The figure illustrates the reconstruction geometry on both horizontal and vertical planes (middle and bottom row). The local coordinate system has its origin lying in the mid-point of the baseline b, (illustrated in figure top-row).

optical-center and landmark center, is ϕ_1 for the left camera and ϕ_2 for the right camera. The camera focal-length is f (in pixels).

The equation in 5.14 computes the landmark z coordinate in the local coordinate system.

$$l_z = \frac{b \cdot \sin(\theta_3) \cdot \sin(\theta_4)}{\sin(\theta_3 + \theta_4)} \tag{5.14}$$

where b and $\vec{l}$ are as defined above. The angles θ_3 and θ_4 are as shown in figure 5.23 bottom-row and can be calculated as:

$$\{ \begin{array}{l} \theta_3 = \arctan(w_v^{left}/f_1) \\ f_1 = \sqrt{f^2 + (w_u^{left})^2} \cdot \cos\theta_1 \end{array} \; ; \quad \{ \begin{array}{l} \theta_4 = \arctan(w_v^{right}/f_2) \\ f_2 = \sqrt{f^2 + (w_u^{right})^2} \cdot \cos\theta_2 \end{array} \; ; \tag{5.15}$$

where f_1 and f_2 are as shown in figure 5.23 bottom-row.

The position of landmark center, $\vec{l}$, is so computed in the local coordinate system. In order to transform the estimated position to the global coordinate system the following equation is applied.

$$\begin{aligned} L_x &= l_x \cdot \cos\alpha - l_y \cdot \sin\alpha + x_{local} \\ L_y &= l_x \cdot \sin\alpha - l_y \cdot \cos\alpha + y_{local} \\ L_z &= l_z + z_{local} \end{aligned} \tag{5.16}$$

where x_{local}, y_{local}, z_{local}, represent the distance between origins of the local and global coordinate systems, (respectively for the X, Y, and Z axis), i.e. the position of the local origin relative to the global system, α is the angle between local and global x axis, and $\vec{L} = (L_x, L_y, L_z)$ is the position of landmark center in the global coordinate system. Figure 5.19 represented these values.

The table 5.1 shows a typical estimation of landmark 3D position based on visual stereo recontruction. The table shows computed 3D positions for the three landmarks represented in figure 5.24. The table also includes ground truth positions (i.e. landmark position measured by the author with a tape meter), and the measured error.

Experiments showed that landmark position can be accurately estimated as long as a match is reliabily found in two panoramic-views. A reliable match can be recognized by a correlation coefficient higher than the established threshold. Experiments have also shown that a reliable match can be easier to be recognized if the camera position is contained into the landmark observation areas (i.e. as defined in subsection 4.1.4), and some ”critical parameters”, (e.g. vergence angle, orientation angle, etc.), as later discussed in section Model Refinement, (5.3), are satisfied. When instead one or more of the critical parameters are above the limit, the matching behavior ”suddenly” changes by not allowing any match to be found anylonger. More details are provide in the following experimentation section.

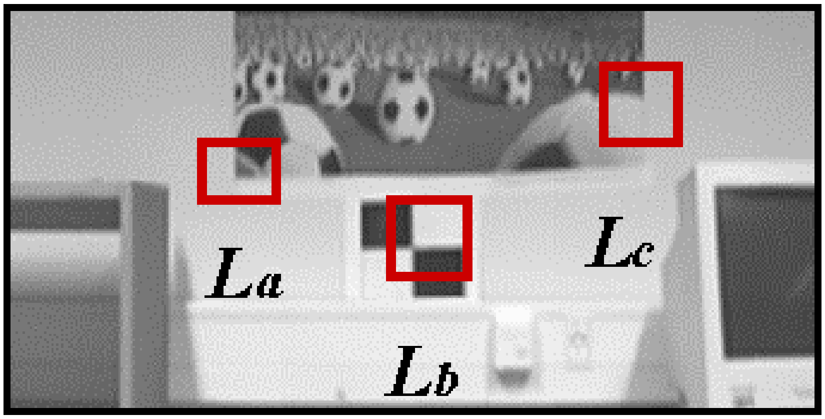


Figure 5.24: The figure shows a wall region of the new laboratory containg a poster and an "ideal" landmark. The landmarks which have been extracted by the proposed system and that have passed the correspondence test are denoted by a red square. The paper representing the "ideal" landmark, L_b, lies 2 mm away from the wall, while the workspace areas corresponding to the selected landmarks L_a and L_c lie 10 mm and 38 mm respectively.

Landmark	cord.	Computed Value (mm)	Ground Truth (mm)	Measured Error (mm)
L_a	x	5783	5804	-21
	y	5568	5542	-26
	z	1126	1108	18
L_b	x	5833	5812	21
	y	4824	4834	-10
	z	1029	1022	7
L_c	x	5764	5776	-12
	y	4515	4516	-1
	z	1099	1247	12

Table 5.1: The table shows estimated 3D position for the landmarks of figure 5.24, (referred to a wall region in the new laboratory).

Landmark Orientation

Unfortunately, to estimate 3D position of landmark center does not provide sufficient information to estimate landmark surface orientation. Consequently, a specific procedure is proposed to estimate the orientation of the surface containing the landmarks. The proposed procedure infers landmark orientation from multiple positional estimates related to the landmark neighbor regions.

The proposed procedure is based on the fact that an image region surrounding a selected landmark and partially containing it, has high probability to be a region with a discriminat texture, and so allowing for a reliable and precise texture match. If the selected landmark is planar, (as required), and the surrounding area allows for reliable matches, the position of the center of the landmark neighbor region can be computed, and this will allow for testing the planarity of the landmark region and estimating the orientation of the surface containing the landmark. In case the planarity condition is confirmed, the landmark orientation can be inferred by estimating the orientation of the plane containing the landmark center.

It is proposed to match landmark neighbor regions using normalized cross correlation and having a template size of the same size as the considered landmark template. This solution is proposed because of the previously discussed advantages of using normalized cross correlation on such a template sizes.

It should be noted that matching landmark neighbor regions requires a significant amount of processing to be done which might considerably slow-down the entire learning process. A limited number of matches would then be advisable to be performed by the system. In particular, it is proposed to only test 4 neighbor regions of a landmark, corresponding to the image regions which have as center a landmark corner (and of the same size of the considered landmark).

The figure 5.25 represents the proposed technique to estimate landmark orientation, (L_θ). In particular, the figure represents the main parameters involved in the equation 5.17, which is proposed to be used to estimate landmark corner position in the case of matching four landmark corner regions. The initial assumption is that all landmarks are verticals.

$$L_\theta = \frac{1}{4}\alpha + \frac{1}{4}\beta + \frac{1}{4}\gamma + \frac{1}{4}\delta \tag{5.17}$$

Figure 5.26 shows then the considered landmark neighboring regions (related to the four corners) for the landmarks a, b and c, represented in figure 5.24.

Note that in principle one landmark corner is sufficient for estimating landmark orientation, whereas additional neighbor matches can be used to confirm the computed landmark orientation. The performance could thus be improved by considering a higher number of neighboring matches. The use of landmark neighbor image-regions which are more distant than those related to landmark corners would also improve accuracy of orientation estimates. However, "distant" regions could represent surfaces which are distinct than landmark region. A solution could then be to test the planarity of a cascade of subsequent neighboring image-regions which would include landmark center and the "distant" neighbor.

The probability for landmark neighbor regions to actually lie on the same surface of the landmark center is much higher when the neighbor region lies in the very proximity of the

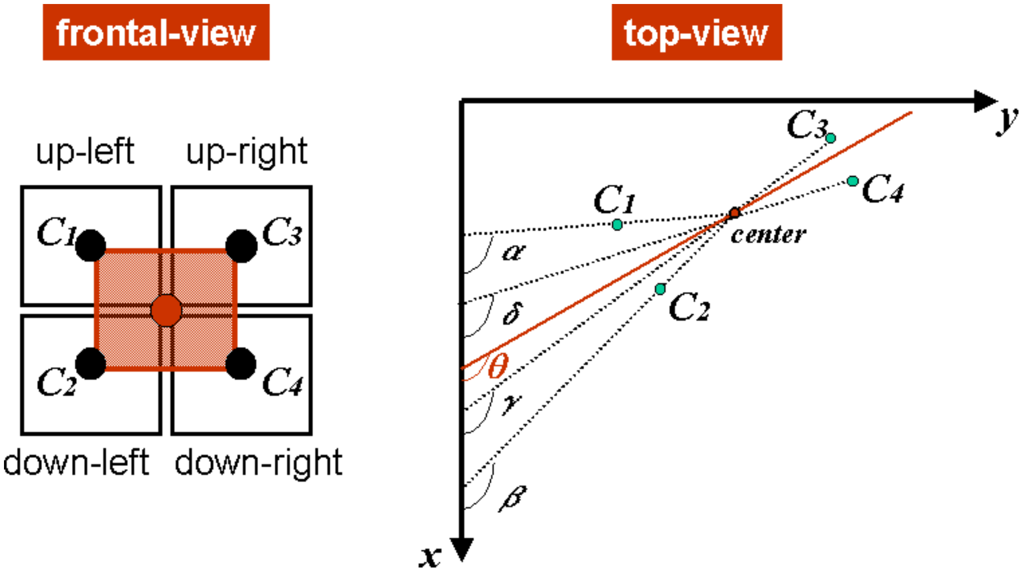


Figure 5.25: The figures represent the proposed technique to estimate landmark orientation, (L_θ), in the case of matching four landmark corner regions. The figure represents the main parameters involved in the equation 5.17 (right-hand), and the landmark and its corner regions (left-hand), surrounded by a red or a black square respectively.

landmark center, (e.g. a distance comparable to the one to landmark corners in our case). However, the closer the region to the center, the larger is the effect of a positional error when inferring landmark orientation.

The accuracy of the computed landmark corner positions plays then a very important role, and it is very relevant to only use the most reliable corner information. This is the reason for proposing the *"ring test"* in order to test reliability of the estimated corner 3D positions.

The "ring test" consists of establishing a volume of the workspace related to each landmark corner expected to contain the computed corner 3D position. The figure 5.27 shows two circular volumes related to landmark corners. The circular volumes can be determined because the system knows the landmark-center 3D position, the size of the texture patch, as well as the camera parameters. A corner which estimated 3D position is not contained in the ring volume (despite its related texture has been matched) is discarded.

Experiments showed the "ring test" to be very relevant in order to achieve an accurate landmark orientation estimate. This in fact allowed for reducing the number of corner information by only keeping the most reliable ones. The main issue in this case is not the quantity but the reliability of the available corner positions.

The experiments showed that the accuracy of the extracted positional information and the constraint provided by the matching. In particular, the orientation can be accurately estimated based on very few positional information. Table 5.2 shows estimated landmark orientation for the extracted texture-patches represented in figure 5.24. The table also includes ground truth and the measured error.

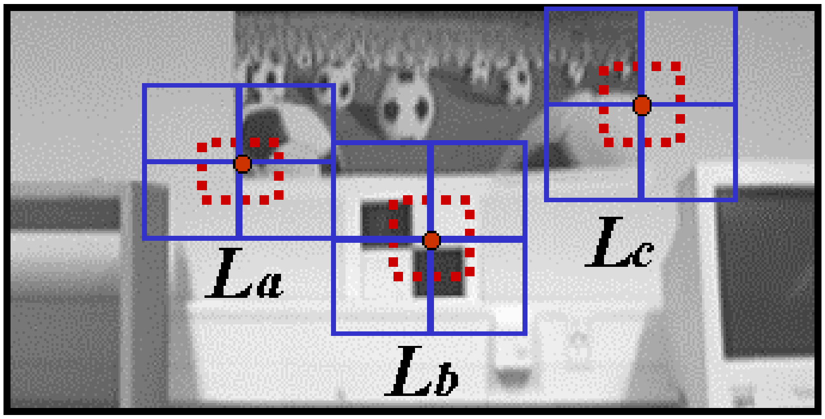


Figure 5.26: The figure shows the considered landmark neighboring regions (related to the four corners) for the landmarks a, b and c, represented in figure 5.24. Landmark regions are in "red" (dashed lines) and corner regions are in "blue".

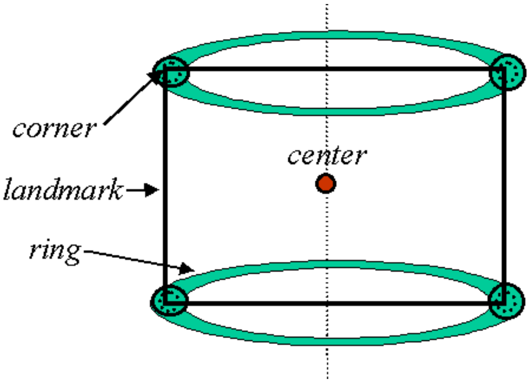


Figure 5.27: The figure shows two circular volumes related to landmark corners representing a volume of the workspace related to each landmark corner expected to contain the computed corner 3D position.

Landmark	orient.	Computed Value (deg)	Ground Truth (deg)	Measured Error (deg)
L_a	γ	94.7	90.0	4.7
L_b	γ	90.9	90.0	0.9
L_c	γ	87.0	90.0	-3.0

Table 5.2: The table shows estimated landmark orientation for the extracted texture-patches represented in figure 5.26. The table also includes ground truth and the measured error.

In order to furtherly improve accuracy of landmark orientation, the error associated with each of the extracted positional information, is proposed to be estimated. Hence, to differently weight the computed position of landmark corners based on the estimated error, when computing landmark orientation. This issue is discussed and demonstrated in section 5.3, (Model Refinement).

5.2.4 Experimentation

The aim of the experimentation was to determine performance of the proposed method for estimating landmark position and orientation in realistic situations. Landmark pose were estimated by stereo reconstruction after extracted texture patches were matched in two panoramic views. The experimentations were performed both in the old and in the new laboratory, but it was only in the new one that landmark orientation was tested.

A rotating panel on a steerable support was introduced in the environment in order to test the orientation of the surface containing the landmark. The panel contained a poster with a quite reach texture. The area surrounding the panel was on the contrary quite plain. Landmarks were consequently expected to represent parts of the panel texture. The panel rotation was only allowed around the vertical axis, and the panel support was specially designed in a way that panel orientation (ground truth) was easy and reliable to determine. Figure 5.29 shows the rotating panel and the steerable support. Figure 5.28 shows panel location in the workspace floor-map of the new-laboratory.

The detected landmarks were mostly lying on walls or on large flat or "almost flat" surfaces, (e.g. on posters which had a few centimeters relief around some of their borders) Some other regions were instead totally planar but not always the area around them, (e.g. the computer monitor and the border regions of the rotating panel). The ideal landmark pattern (figure 5.29 center) were also introduced to test performance discrepancies between ideals and naturally occurring patterns.

During the experimentation the robot was taken to the 7 workspace locations represented in figures 5.28 top and bottom rows, (i.e. the same positions already illustrated in figures 5.13 and 5.14). The different locations led to different distances and orientations to landmarks, and to different baseline configurations, i.e *A-baseline* $= (A_l - A_r)$, *B-baseline* $= (B_l - B_r)$, *C-baseline* $= (C - B_r)$, and *LR-baseline* $= (L - R)$.

The camera position was estimated for each robot location by means of a specially designed procedure. The proposed procedure was based on triangulation of three *specialized landmarks* which were introduced in the laboratory. Figure 5.29 right-hand shows typical specialized landmarks (little squares). In particular, by using the computer mouse to manually point to the image-pixel corresponding to the center of each specialized landmark, the correspondent landmark image-locations could easily be extracted and then provided as input to a program which used triangulation to compute camera position and orientation in the workspace. This result could be achieved since the position of the specialized landmarks in the workspace and camera focal length were known a priori.

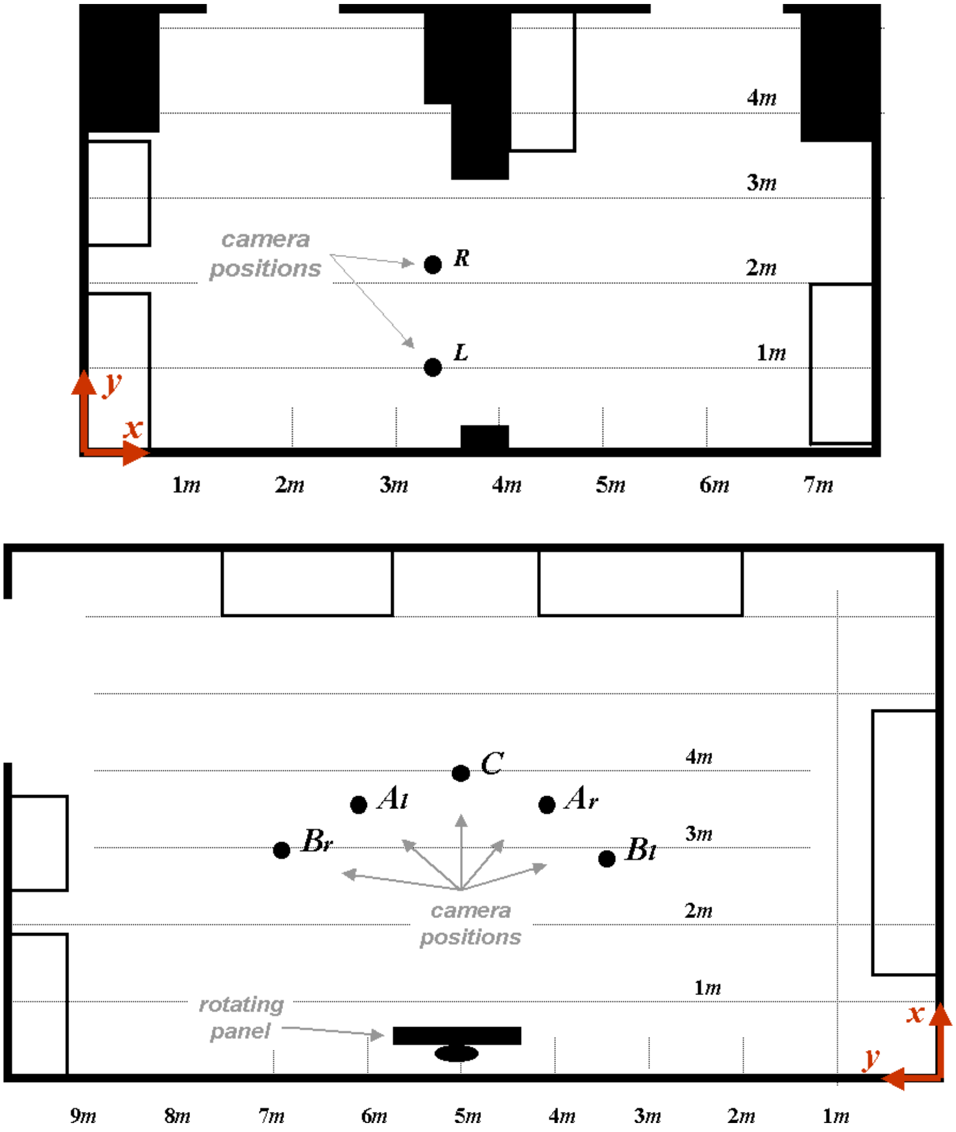


Figure 5.28: The figures represent floor-maps of the old (top-row) and the new (bottom-row) laboratories. The figures show the position of the rotating panel and the 7 workspace locations where the robot was taken to acquire workspace views, (the same positions have been already illustrated in figures 5.13 and 5.14).

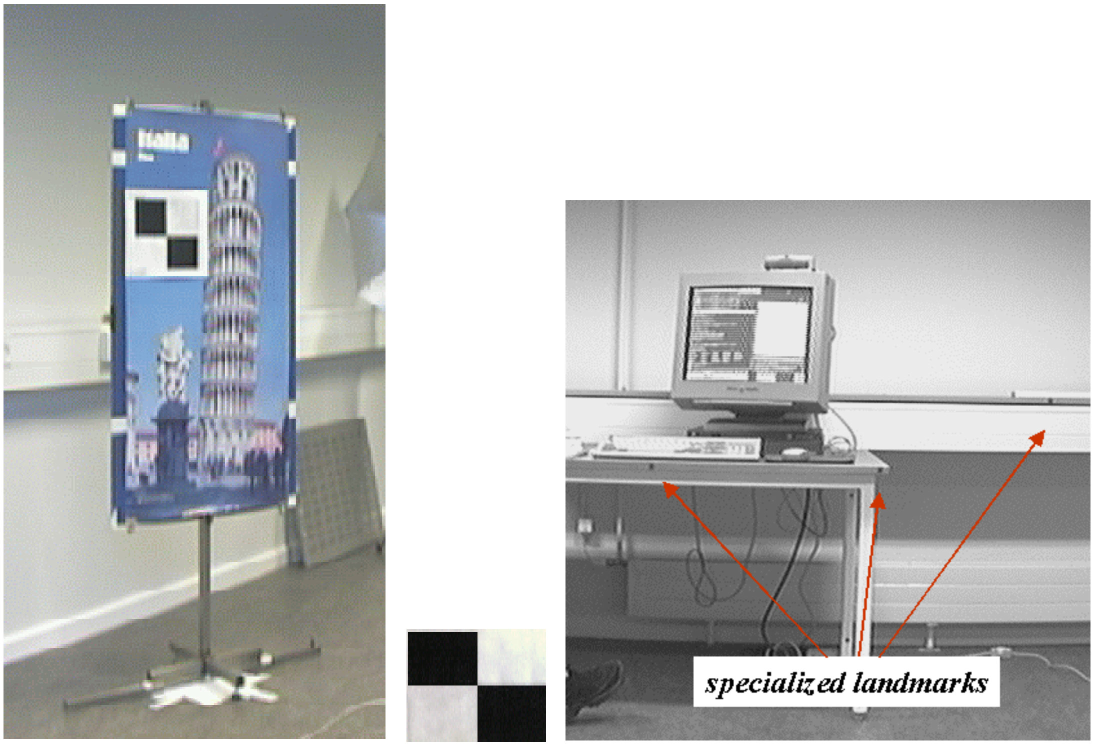


Figure 5.29: The figure left-hand side shows the rotating panel on the steerable support, and the figure in the center shows the "ideal" landmark, both introduced in the environment to better test system performance. The figure right-hand shows a typical pattern for the specialized landmarks (little squares), and their position in the environment.

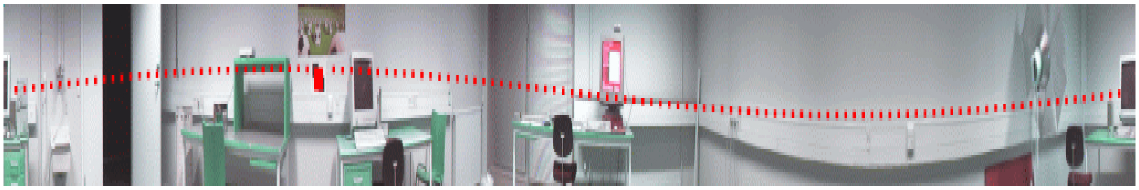


Figure 5.30: The figure represents the cylindrical epipolar line for the landmark L_b of figure 5.24 superimposed to the associated panoramic view (from position A_r in the new laboratory). The figure also represents the identified match, (the red square).

The proposed technique based on three specialized landmarks allowed for a very accurate camera pose estimate, (average error around 1 cm.), when compared with that measured by hand. It was estimated that the real camera pose can be measured by hand with an accuracy of approximately 2 cm. A lower positional accuracy in camera position, (average error 3 cm), was also simulated during the experiments in order to test the system with a camera positional error representing the one obtained in the localization experimentation (see sub-section 4.4), where the robot was let to move autonomously under the control of the proposed self-localization algorithm, (Livatino and Madsen [93]).

At each position the robot stopped and ran the procedures described in this chapter concerning synthesis of panoramic views and extraction of landmark representative images. Then, the proposed algorithms for landmark pose computation were run, and achieved results compared to ground truth. The landmark pose ground-truth was measured by hand using a tape meter with an accuracy of approx. 1 cm. It can be noticed that due to the complexity of the robotic system, it was easier, and thus more precise, to measure ground truth related to landmark poses (1 cm. average error), than to camera poses (2 cm. average error).

Correspondences Test

A panoramic view was synthesized for each of the 7 robot positions represented in figures 5.28. Matches were then searched on couple of correspondent panoramic-views (i.e. those forming baselines A, B, C, and LR). Extracting landmarks on different panoramas often generated different responses even if the views concerned the same workspace region. This was mainly due to different distortions arising from different viewpoints, as well as to the size of the generated landmark texture patches.

As previously mentioned, it was decided that for each pair of available panoramic views, texture patches were extracted in the left panoramas and correspondences searched in the right panoramas. The process could additionally be run on the opposite way. Nevertheless, for the purpose of the thesis, which concerned with testing the system, this was considered not necessary to implement.

The process of landmark extraction was tested many times with different parameter values. This led to different subsets of extracted landmarks which showed a comparable level of discriminance and invariance. A higher number of landmarks were selected when the attentive scheme was "tuned" on the type of environment texture. The advantage of running the attentive process many times with different parameters is to obtain a larger set of candidate landmarks.

The search for correspondences along cylindrical epipolar lines (using normalized cross correlation) was implemented according to the method presented in subsection "Correspondence Test", (5.2.2). The experimentation related to correspondences test concerned reliability of computed template matching and, in particular, the performance in relation to the associated correlation coefficient. An extensive number of runs were then performed in order to test possible and typical responses to normalized cross correlation while searching along epipolar lines. The reliability of each match was measured by analyzing the resulting correlation coefficient associated to each successful or misidentified match.

Figure 5.30 represents the epipolar line for the landmark L_b of figure 5.24 superimposed to the associated panoramic view (from position A_r in the new laboratory). The figure also represents the identified match.

Diagrams in figure 5.31 show typical correlation responses when searching for matches along epipolar lines. In particular, the figure shows diagrams representing the resulting correlation coefficient associated to different columns in the considered sub-panoramic view (i.e. the image represented in figure 5.24). The value of the column is derived from the angle γ showed in equation 5.10. The first number in the diagram abscissas represents the image-column correspondent to the center of the search window, while the second number (in brackets) represents the column value correspondent to the highest correlation value (resulting from the match). The threshold value for the resulting correlation coefficient was set to 0.75. It can be noted that in case of valid matches, the same image location was confirmed in two subsequent search windows (each two subsequent search windows always intersect each other).

It was experienced that the correlation coefficient gets a low value (below 0.4) when the match response is unreliable. On the other hand, it gets a high value (above 0.7) both on valid matches and on some false matches. In case of false matches (and high value), the matches were mostly related to patches which had low-contrasted textures. The correlation coefficient, consequently, can not be the only value we should look at to be sure that we are in presence of a valid match. Nevertheless, if low contrasted patches could in some way be discarded (for example, by further tests), a high correlation value would only mean a valid match.

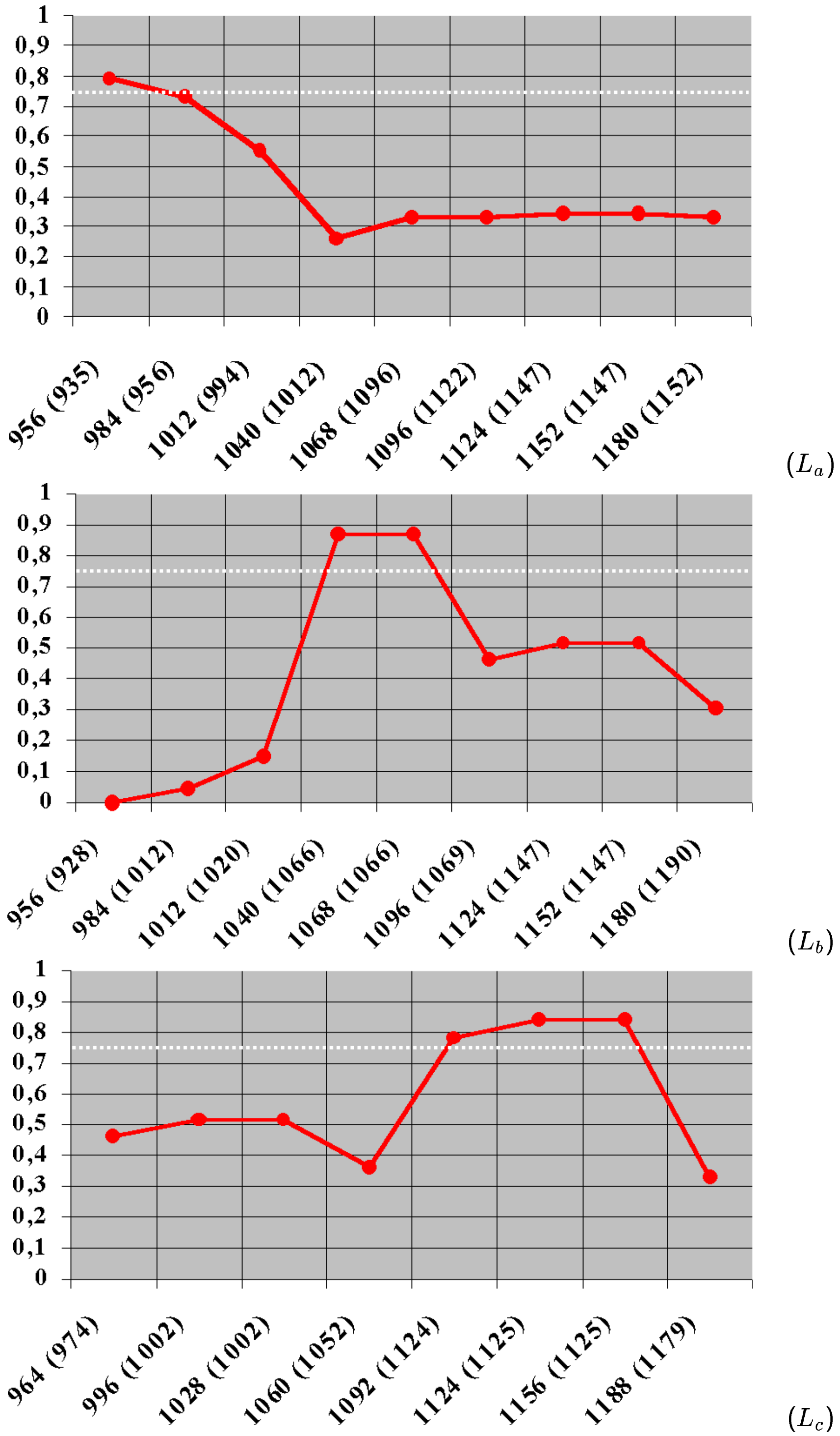
The proposed technique of searching for matches along cylindrical epipolar lines demonstrated to be a necessary step which had not only the role of reducing processing time (by limiting the searching area), but it also dramatically reduced the number of false matches. In particular, all those matches which would have taken place outside the epipolar line search region.

Interestingly, it was also experienced that most of the low-contrasted textures which generated high correlation coefficients when matched, led to generation of *multiple matches*[2], if the searching algorithm was kept running, (so looking for further matches), along the epipolar line. Multiple matches were on the contrary not found in cases of valid-matches with high-value correlation coefficient.

The proposed algorithm for correspondences test was consequently required to scan the entire epipolar line before that a positive response (valid match) could be provided. Indeed, there is not a real need to scan the entire epipolar line to confirm a valid match. In fact, if a high correlation value corresponds to a false match, it will very shortly generate a multiple match, so that scanning a "limited" portion of the epipolar line can be sufficient. In our experiments sub-panoramic regions of 512 columns were considered. The proposed procedure was thus let to run until either a multiple matched was found or the end of the epipolar-line portion was reached.

An interesting issue was to experience that searching for multiple matches also allowed to identify, and thus discard, extracted texture patches which, despite they showed high-contrasted textures, sharp-edges, and local symmetries, they could easily become ambiguous.

[2] *Multiple matches* refers to the case when different image regions match the same template.



(L_a)

(L_b)

(L_c)

Figure 5.31: The diagrams in figures show typical correlation responses when searching for matches along epipolar lines (see text).

Figure 5.32 shows an example situation where from the 15 patches extracted by running the attentive scheme described in subsection 5.1.3, (figure top-row), only 3 of them were selected after the correspondences test (figure bottom-row). In particular, the example in the figure showed landmarks which were discarded because of:

- Low-contrasted textures, (patches e, g, r). These patches easily generate multiple matches.
- Ambiguous textures, (patches d, f, i, l). These patches easily generate multiple matches.
- Different visible aspects, (patches m, n, o, p). A match can not be found or the normalized cross-correlation algorithm generates a low correlation coefficient on the region correspondent to the extracted patch.
- Occlusions / disocclusions (and shadows) have arisen, (patches e, h). A match can not be found, or the normalized cross-correlation algorithm generates a low correlation coefficient on the region correspondent to the extracted patch.

The only patches which went through the test are those labeled a, b, c in figure 5.32. They in fact generated a unique image region matched with a high correlation value.

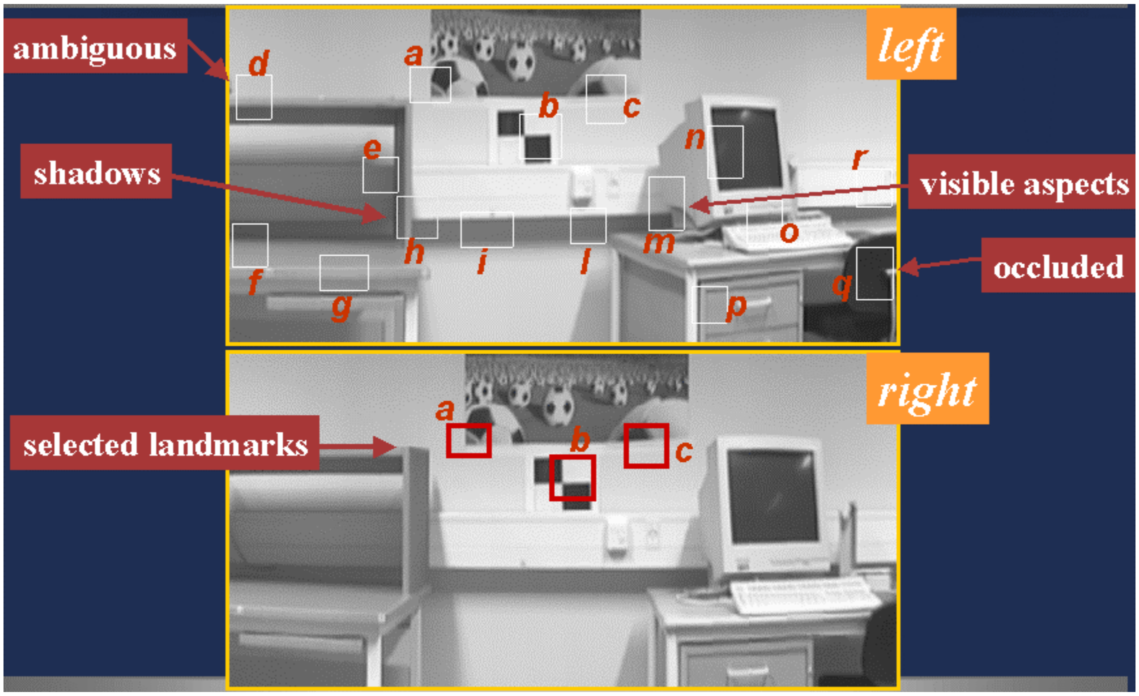


Figure 5.32: The figure shows an example of landmarks rejected while testing for correspondences (see text). Landmarks were discarded because they represented texture low-contrasted, ambiguous, with different visible aspects, containing occlusions/disocclusions (and shadows), etc., (see text).

Response	Total Number	match (num)	Average CC	Max Val CC	Min Val CC
Correct	744	yes (93)	0.80	0.93	0.56
	(93%)	not (651)	0.35	0.54	0.12
False	56	yes (48)	0.91	0.96	0.63
	(7%)	not (8)	0.52	0.549	0.49

Table 5.3: The table represent a "statistic" based on 800 matches which involved the correspondence test with the different analyzed environment views. "CC" stands for correlation coefficient.

A "statistic" based on 800 matches which involved the correspondence test with the different analyzed environment views, (more environment views as shown later in further experimentation), are summarized in table 5.3.

The lesson learned from this experimentation was:

1. Main cause of false matches with a high correlation coefficient are low-contrasted texture-patterns.
2. Ambiguous matches as well as low-contrasted ones, may be discarded by searching for multiple matches.
3. The normalized cross correlation allows the system to reject landmarks observed with a high perspective distortion, landmarks showing different visible aspects of the same object, and landmarks affected by undesired illumination effects such as reflections, highlights and shadows, or which easily become occluded.
4. It is very important with a correlation threshold which is tuned on the environment characteristic. In the performed experiments, a threshold of 0.55 was experienced as the optimal threshold value, (i.e. minimizing the number of false matches).
5. Still, around 7% of the false matches were able to pass through the correspondences test.

Landmark Position

The position of environment characteristics corresponding to matched texture patches was estimated for each patch which went through the correspondences test using equations 5.12 and 5.14, and then 5.16 to convert to global frame (see sub-section 5.2.3). The result of computed 3D positions were related to landmark centers. The estimated values were then compared to ground truth to measure their accuracy.

Table 5.1 showed estimated 3D positions for the landmarks of figure 5.24, (a wall region in the new laboratory), and table 5.4 shows a typical result associated with the landmark showed in figure 5.33, (a more cluttered wall region in the old laboratory). The tables also include ground truth. The distance between the camera centers, the baseline, was 0.97 meters in case of the example of table 5.1 (*A-baseline*), and 1.19 meters in case of the example of table 5.4 (*LR-baseline*).

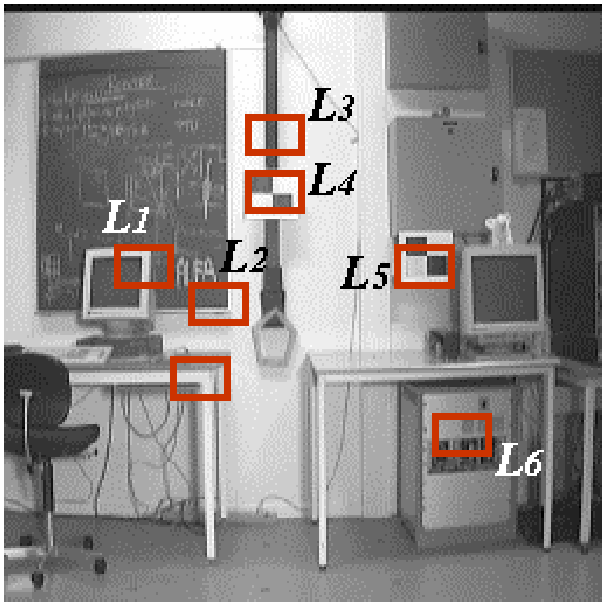


Figure 5.33: The figure shows a region of the old laboratory containing several objects and two "ideal" landmarks. The landmarks which have been extracted by the proposed system and that have passed the correspondence test are denoted by a red square with an associate number.

Landmark	cord.	Computed Value (mm)	Ground Truth (mm)	Measured Error (mm)
L_1	x	614	559	55
	y	1135	1158	-23
	z	1118	1101	17
L_2	x	88	50	33
	y	1869	1860	9
	z	916	923	-7
L_3	x	186	135	51
	y	2350	2370	-20
	z	1739	1725	14
L_4	x	91	135	-44
	y	2357	2370	-13
	z	1493	1475	18
L_5	x	424	386	38
	y	2982	2970	12
	z	1188	1205	-17
L_6	x	661	614	47
	y	3133	3100	33
	z	410	430	-20

Table 5.4: The table shows estimated 3D position for the landmarks of figure 5.33, (referred to a wall region in the old laboratory).

Tables 5.5 and 5.6 show position estimates for the landmarks detected in the rotating panel. The baseline were *C-baseline* and the *B-baseline* respectively. Additional results are later shown in chapter 6, (tables 6.1 and 6.2 in sub-section 6.4.1), where the system capability of automatically learn landmarks will furtherly be tested in the context of automatic recognition of self-learned landmarks.

The obtained results were very encouraging and demonstrated:

- The accuracy of the proposed method. For each correct landmark match, (i.e. having measured error in image-location below 3 pixels), the error in computed landmark 3D positions ranged from 0 through 7.2 cm., (for the worst performing coordinate axis), with 3 cm. average error.
- The higher precision when using a larger baseline. For example, when comparing tables 5.5 and 5.6, (the latter related to a larger baseline configuration), a general improvement in estimated accuracy can be noted.
- The reliability of the proposed method. The 93% of the experimented cases represented correct matches, (see table 5.3).
- The fundamental role of the correspondences tests. In fact, landmarks which went through the correspondences test produced for most of the cases (93%) accurate position estimates.

A very interesting outcome of the experiments was to experience that in almost all the cases where landmark position was not correctly computed despite a successful correspondences test, the estimated position was "largely" wrong, i.e. over 15 cm error in at least one of the coordinate axis. The positive aspect of having a large error is that it has high probability to be recognized in further "consistency tests". For example, when estimating the landmark surface orientation, and in case of additional knowledge, such as an environment map, is available.

The accuracy of estimated landmark position was also tested in relation to the *"orientation angle"*, (i.e. the angle between landmark surface and baseline). This topic has already been introduced in sub-section 4.1, (concerning observation areas), and it will thoroughly be discussed in chapter 6. As for now, we can summarize that the accuracy of landmark position estimate remains basically unchanged until when a critical value for the "orientation angle" is reached (that is, around 30 degrees). Beyond this critical value landmark position can not be estimated because no texture correspondence can usually be found. This behavior, which can be called "on-off", will be described in chapter 6 (figure 6.22), where the value of correlation coefficient is reported according to different landmark positions and orientations to a landmark.

Landmark Orientation

In order to estimate landmark surface orientation the technique presented in the above method description (sub-section 5.2.3) was implemented and tested on different landmarks. The proposed technique required that four additional image-templates related to corner regions, of the same size of the considered landmark, would be matched by using the normalized cross-correlation.

The table 5.2 showed typical landmark orientations estimates based on visual stereo reconstruction, and the figure 5.26 showed the position of the corner from where the orientation of the landmark surface was estimated.

As expected, the result of the template matching related to landmark corners was as accurate as for the landmark center. Though, the corner regions were not always matched. In particular, the experiments showed that in many situations it is not possible to match all landmark corners since these may not represent a planar surface or they have low-contrasted or ambiguous textures. In these cases, the system must rely on fewer corner positions. Nevertheless, the reliability of the matched information and the strong constraint provided by the geometric constraints, demonstrated to be sufficient for an accurate orientation estimate even in case of only one matched corner position, i.e. the minimum number.

In case the system matches more than one corner region, the multiple available corner information can be used to refine and make more reliable the estimate of landmark orientation. In particular, the estimate will be more accurate because each corner related orientation can be averaged with the others. More reliable because multiple "similar" orientation information validate the estimated orientation.

Figure 5.34 shows which of the corner regions represented in figure 5.26 allowed for a match, hence, to estimate the landmark orientation value reported in table 5.2.

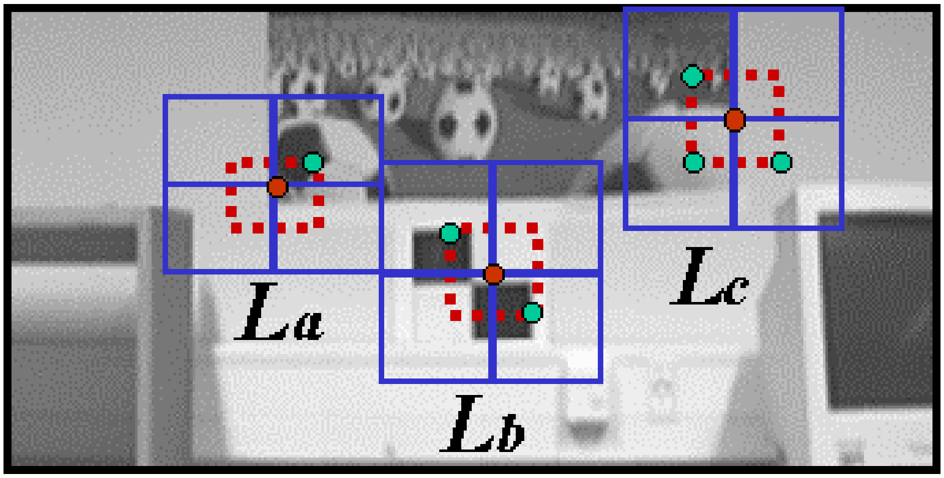


Figure 5.34: The figure shows which of the corner regions represented in figure 5.26 allowed for a match, (regions with a green "dot"), hence, to estimate the landmark orientation values reported in table 5.2.

As can be noted in figure 5.34 there was only one corner match for landmark L_a, the other corner regions were not matched because they had either a low contrasted texture (region top-left and bottom-right) or different visible aspects (region ”bottom-left”). In case of landmark L_b, the reason for the unmatched corner regions (”top-right” and ”bottom-left”) is mainly due to ambiguous textures. In case of landmark L_c, the reason for the unmatched corner region (”top-right”) is a low contrasted texture. Note that in case of landmark L_c, the positional information related to the corner region ”bottom-right” was not considered to estimate landmark orientation despite the patch was matched, because the extracted 3D point was not lying in the plane passing through the landmark center and the other two corners. This situation can also visually be noted: the corner region includes a portion of a computer monitor which lies in another plane than the poster.

In further experiments the camera system was set to allow for detection of landmarks representing portion of a poster hanging on the rotating panel shown in figure 5.29 left-hand. In this way, computed landmark surface orientation could easily be compared to ground truth and tested for different angles. The table in 5.5 shows computed orientation for landmark of figures 5.35. The distance between the cameras, (the baseline), was in this case case 1.35 meters (*C-baseline*).

The experiments showed the accuracy provided by the proposed method. This is also based on the fact that position of landmark center and corners can precisely be estimated (with few centimeters average error), if matched. Hence, if a landmark possesses a planar neighborhood which can reliably be matched, the estimated corner positions pass the ”ring test”, and the final estimate is based on an averaged sum, the resulting landmark orientation angle has a high probability to represent an accurate and reliable estimate. On the other hand, if a selected landmark does not posses a planar neighbor, or it does not allow for a match, the landmark orientation can not be inferred and the landmark will be discarded.

A critical factor is the size set for the ”ring” volume in the ”ring test”, as well as the tolerated discrepancy between different angle estimates associated to each corner, (in the example of sub-section 5.2.3, the discrepancy among the orientation estimates α, β, γ, and δ, was depicted in figure 5.25). Further experimentation, (section 5.3) will show that landmark accuracy is definitely improved by weighting lanmark corner positions on their expected precision, (where the expected precision is based on the analysis of the error propagation).

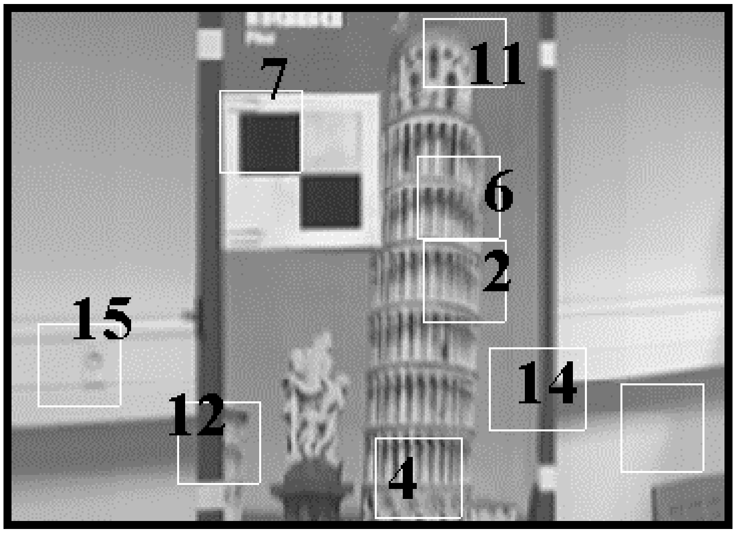


Figure 5.35: The figure shows a region of the new laboratory containing the rotating panel and an "ideal" landmarks. The landmarks which have been extracted by the proposed system and that have passed the correspondence test are denoted by a white square with an associate number.

Landmark	cord.	Computed Value (mm)	Ground Truth (mm)	Measured Error (mm)
L_2	x	849	802	47
	y	5036	5056	-20
	z	1010	997	13
	γ	73.3 (deg.)	70.0 (deg.)	3.3 (deg.)
L_4	x	812	772	40
	y	5009	4960	49
	z	754	713	41
	γ	63.0 (deg.)	70.0 (deg.)	-7.0 (deg.)
L_6	x	858	802	56
	y	5017	5042	-25
	z	1127	1107	20
	γ	71.8 (deg.)	70.0 (deg.)	1.8 (deg.)
L_7	x	582	622	-40
	y	4656	4648	8
	z	1207	1195	12
	γ	68.9 (deg.)	70.0 (deg.)	-1.1 (deg.)
L_{11}	x	852	802	50
	y	5067	5050	17
	z	1354	1333	21
	γ	72.5 (deg.)	70.0 (deg.)	2.5 (deg.)

Table 5.5: The table shows estimated 3D position and orientation for the landmarks of figure 5.35.

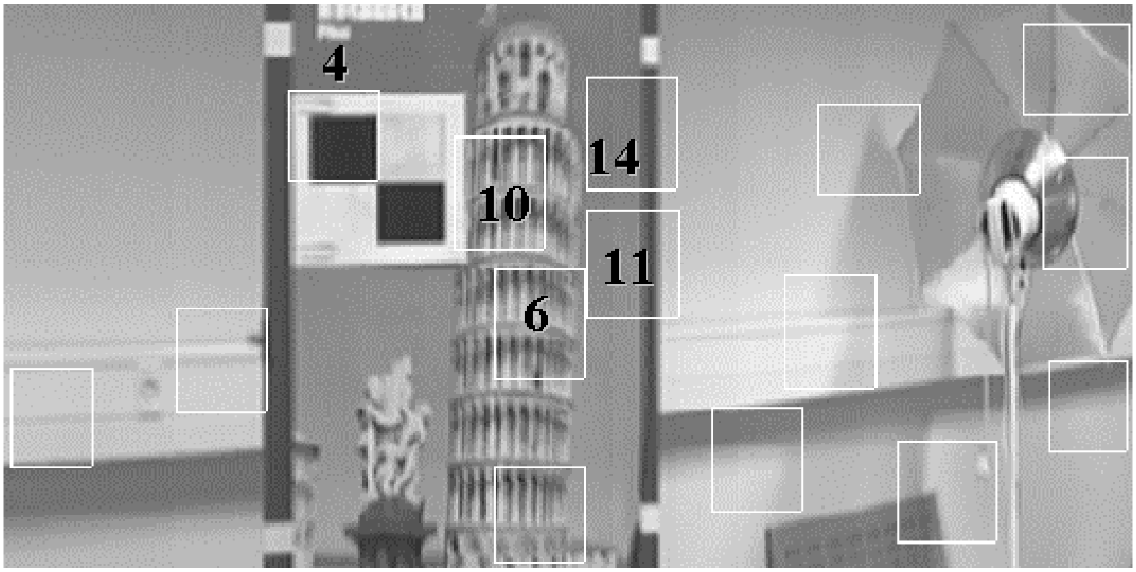


Figure 5.36: The figure shows a region of the new laboratory containing the rotating panel and an "ideal" landmarks. The landmarks which have been extracted in a typical run by the proposed system and that have passed the correspondence test are denoted by a white square with an associate number.

Landmark	cord.	Computed Value (mm)	Ground Truth (mm)	Measured Error (mm)
L_4	x	832	802	32
	y	5020	4980	40
	z	1018	997	21
	γ	64.9 (deg.)	70.0 (deg.)	-5.1 (deg.)
L_6	x	806	772	34
	y	4899	4911	-12
	z	723	713	10
	γ	68.0 (deg.)	70.0 (deg.)	-2.0 (deg.)
L_{10}	x	776	802	-26
	y	4967	4974	-7
	z	1117	1107	10
	γ	69.5 (deg.)	70.0 (deg.)	-0.5 (deg.)
L_{11}	x	601	622	-21
	y	4657	4648	9
	z	1187	1195	-8
	γ	70.1 (deg.)	70.0 (deg.)	0.1 (deg.)
L_{14}	x	752	802	-50
	y	4939	4980	-41
	z	1313	1333	-20
	γ	84.9 (deg.)	70.0 (deg.)	14.9 (deg.)

Table 5.6: The table shows estimated 3D position and orientation for the landmarks of figure 5.36.

5.3 Model Refinement

The last step of the learning process is for refining the model built at previous steps. This step is proposed because there still is a certain amount of learned information, either wrong or imprecise, that the landmark extraction process and the correspondence test are not able to ”filter”. A good part of this wrong or imprecise information can nevertheless be revealed through a series of ”consistency tests” based on the learned positional information, and through the estimation of landmark positional uncertainty.

The landmarks which go through all tests during the model refinement step are *elected* to represent the learned environment model, thus they are stored into the system database to be used during the self-localization phase to estimate robot pose.

The first sub-section (5.3.1) deals with the issue of uncertainty characterization. The aim is to discard the most uncertain information, and to classify the ones kept, based on the amount of the estimated covariance values. The second sub-section (5.3.2) deals with testing consistency of the information kept through the analyisis of two critical parameters: the vergence and the orientation angle. The aim is to discard ”inconsistent” information and to *elect* the remaining ones.

5.3.1 Uncertainty Characterization

The error in extracted landmark pose has a large impact on robot localization performance. The three main reasons are:

a. An error in the position of computed landmark center and neighbor regions affects the accuracy of landmark orientation estimate. The experiments later in this section demonstrate this concept by showing that a more reliable orientation estimate can be provided if the system is able to estimate the error associated with the landmark positions.

b. An error in landmark pose affects the accuracy of robot pose estimate by playing a role in the triplet selection process, (sub-section 4.2.3). In particular, a more precise robot localization can be performed if the triplet-selection process also takes into account the uncertainty associated to landmark poses and the distribution of such uncertainty. The figure 5.37 represents how robot positional uncertainty varies when landmarks have a different uncertainty distribution.

c. An error in landmark pose affects the accuracy of robot pose estimate by playing a role in the triangulation-based robot localization process, (sub-section 4.2.1). In particular, the smaller the error in landmark pose the more accurate and reliable is the robot pose estimate (and the associated estimated uncertainty).

In summary, the proposed uncertainty characterization would be beneficial to the system in terms of: a more precise landmark orientation estimate (case *a*), a more appropriate choise of landmark triplets (case *b*), and a more accurate robot pose estimate (case *c*).

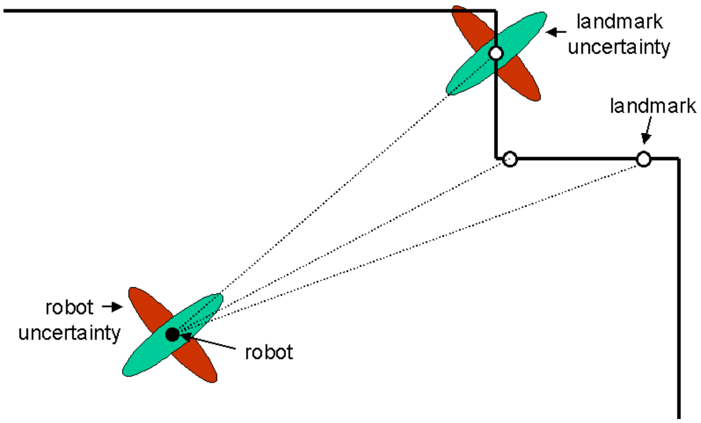


Figure 5.37: The figure describes a portion of an example workspace floor-map. The figure shows how in principle robot positional uncertainty varies when landmarks have a different uncertainty distribution. In the represented case, because of the equation used by the triangulation, the error in the "left landmark" is directly transfered to robot position (in case there is not error for the other landmarks). The red ellipse associated to current robot position represents the uncertainty on the robot position resulting when the uncertainty in the left landmarks is the red one. Analogous the case of the green ellipse.

Sensitivity Analysis

Landmark 3D position is inferred by the stereo situation arising when the same workspace area is observed from different view points. Since the robot has only one camera, the views from at least two different robot positions are required, which leads to the fact that the robot is required to move during the learning phase, (see figure 5.23 top-row). Consequently, it is important to consider the error associated with robot pose, which in turn leads to an error in camera pose, during the automatic learning.

It has to be noticed that the error affecting camera pose is not the only cause of uncertainty in landmark position. In fact, as demonstrated in Madsen and Andersen [97] and discussed in sub-section 4.2.2, there can also be an error in the landmark image-locations when these are detected by image correlation.

It is then important to understand how an error in the input parameters, i.e. camera pose and landmark image-location, affects landmark positions estimates when this is computed based on stereo triangulation. In other words, what is the sensitivity of the output parameters towards noise in the input parameters?

The technique for sensitivity analysis previously described in chaper 4 (sub-section 4.2.2), represents a suitable way of estimating propagation of errors in the input parameters, thus, this technique is proposed to estimate landmark positional uncertainty.

Since the robot is moving on a planar surface, the z component of camera position is not considered when computing the error propagation. As for the camera focal-length, this is considered very precisely estimated by an off-line process (see 4.1.5), therefore, is not considered. Consequently, the input variables of interest when computing landmark position by the proposed stereo reconstruction can be described by the vector

$\vec{k} = [c_x^{left}, c_y^{left}, c_x^{right}, c_y^{right}, w_u^{left}, w_v^{left}, w_u^{right}, w_v^{right}]$, where (c_x^{left}, c_y^{left}) and $(c_x^{right}, c_y^{right})$ respectively represent the camera "left" and "right" position in the local coordinate frame, and (w_u^{left}, w_v^{left}), and $(w_u^{right}, w_v^{right})$ respectively represent the landmark image location in the "left" and "right" camera image-plane.

If we call g the function used to compute landmark position (as defined in equations 5.12 and 5.14), the resulting landmark position, i.e. $\vec{l} = [l_x, l_y, l_y]$, can be obtained by the equation:

$$\vec{l} = g\,(\,\vec{k}\,) \tag{5.18}$$

The equation that in term of one-dimensional example was described as $\Delta y = \Delta x \cdot g'(x_0)$ for the first order approximation, can thus be described in terms of multi-dimensions as in the following:

$$\Upsilon \;=\; \frac{\partial \vec{l}}{\partial \vec{k}} \;\Lambda\; \frac{\partial \vec{l}}{\partial \vec{k}}^{T} \tag{5.19}$$

where Υ is the covariance in the input parameters. Assuming all the input parameters in $\vec{k}$ independents, Λ can be described as in the following:

$$\Lambda = \begin{bmatrix} \sigma^2_{c_x^{left}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{c_y^{left}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{c_x^{right}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{c_y^{right}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{w_x^{left}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{w_y^{left}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_{w_x^{right}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_{w_x^{right}} \end{bmatrix} \tag{5.20}$$

where the matrix of the partial derivatives $\frac{\partial \vec{l}}{\partial \vec{k}}$, is described by:

$$\frac{\partial \vec{l}}{\partial \vec{k}} = \begin{bmatrix} \frac{\partial l_x}{\partial c_x^{left}} & \frac{\partial l_x}{\partial c_y^{left}} & \frac{\partial l_x}{\partial c_x^{right}} & \frac{\partial l_x}{\partial c_y^{right}} & \frac{\partial l_x}{\partial w_u^{left}} & \frac{\partial l_x}{\partial w_v^{left}} & \frac{\partial l_x}{\partial w_u^{right}} & \frac{\partial l_x}{\partial w_v^{right}} \\ \frac{\partial l_y}{\partial c_x^{left}} & \frac{\partial l_y}{\partial c_y^{left}} & \frac{\partial l_y}{\partial c_x^{right}} & \frac{\partial l_y}{\partial c_y^{right}} & \frac{\partial l_y}{\partial w_u^{left}} & \frac{\partial l_y}{\partial w_v^{left}} & \frac{\partial l_y}{\partial w_u^{right}} & \frac{\partial l_y}{\partial w_v^{right}} \\ \frac{\partial l_z}{\partial c_x^{left}} & \frac{\partial l_z}{\partial c_y^{left}} & \frac{\partial l_z}{\partial c_x^{right}} & \frac{\partial l_z}{\partial c_y^{right}} & \frac{\partial l_z}{\partial w_u^{left}} & \frac{\partial l_z}{\partial w_v^{left}} & \frac{\partial l_z}{\partial w_u^{right}} & \frac{\partial l_z}{\partial w_v^{right}} \end{bmatrix} \tag{5.21}$$

The resulting covariance matrix will then be as in the following:

$$\Upsilon = \begin{bmatrix} \sigma_{l_x}\sigma_{l_x} & \sigma_{l_x}\sigma_{l_y} & \sigma_{l_x}\sigma_{l_z} \\ \sigma_{l_y}\sigma_{l_x} & \sigma_{l_y}\sigma_{l_y} & \sigma_{l_y}\sigma_{l_z} \\ \sigma_{l_z}\sigma_{l_x} & \sigma_{l_z}\sigma_{l_y} & \sigma_{l_z}\sigma_{l_z} \end{bmatrix} \tag{5.22}$$

In order to represent the computed positional uncertainty in the global coordinate system, Υ must be rotated according to the rotation matrix Rot_z, where α is the angle between local and global x axis.

$$Rot_z = \begin{bmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.23}$$

Typical covariance matrices are shown in the followings lines. They refer to the landmarks L_4, L_6, L_{11}, L_{10}, and L_{14}, depicted in figure 5.39, (the same as in figure 5.36). The Υ_{L_i} represents the covariance matrix of landmark L_i, with $i = 4, 6, 10, 11, 14$, (values in squared millimeters), and $\sigma^2_{L_\theta}$ represents the landmark orientation variance, (values in squared degrees).

$$\Upsilon_{L_4} = \begin{bmatrix} 5513.4 & 233.6 & 56.2 \\ 233.6 & 38.5 & 9.3 \\ 56.2 & 9.3 & 25.3 \end{bmatrix} \qquad \sigma^2_{L_\theta} = 41.2712 \tag{5.24}$$

$$\Upsilon_{L_6} = \begin{bmatrix} 5286.2 & 346.5 & 60.7 \\ 346.5 & 25.3 & 4.0 \\ 60.7 & 4.0 & 9.4 \end{bmatrix} \qquad \sigma^2_{L_\theta} = 53.6526 \tag{5.25}$$

$$\Upsilon_{L_{10}} = \begin{bmatrix} 5285.1 & -881.4 & 93.5 \\ -881.4 & 8.98 & 15.3 \\ 93.5 & 15.3 & 20.1 \end{bmatrix} \qquad \sigma^2_{L_\theta} = 0.02655 \tag{5.26}$$

$$\Upsilon_{L_{11}} = \begin{bmatrix} 5285.7 & 17.0 & 4.1 \\ 17.0 & 2.8 & 0.03 \\ 4.1 & 0.03 & 4.7 \end{bmatrix} \qquad \sigma^2_{L_\theta} = 0.0227 \tag{5.27}$$

$$\Upsilon_{L_{14}} = \begin{bmatrix} 5353.5 & 753.1 & -114.4 \\ 753.1 & 109.1 & -16.3 \\ -114.4 & -16.3 & 29.5 \end{bmatrix} \qquad \sigma^2_{L_\theta} = 64.5817 \tag{5.28}$$

Table 5.7 reports the orientation estimates related to landmarks represented in figure 5.39. In particular, table 5.7 shows the orientation values calculated on the basis of matched landmark corners information. The matched landmark corners are also illustrated figure 5.39 (green circles). The figure 5.38 (the same as figure 5.25) reminds us about the correspondence between number and corner position relative to landmark center.

Equation 5.29 describes how to compute landmark orientation. This equation is derived from equation 5.17, but it is this time weighted on the associate covariance values. The weights k_i in the equation, with $i = 1, 2, 3, 4$, are calculated based on how many corner information have been matched and have passed the "ring test".

$$L_\theta = k_1\ \alpha + k_2\ \beta + k_3\ \gamma + k_4\ \delta \tag{5.29}$$

The landmark orientation is more accurate when the positional information related to corner positions and landmark center are weighted by their expected uncertainty. This can be noted by comparing table 5.7 with table 5.6. To help "seeing" this difference, the landmark orientations of table 5.6 are reported in table 5.7 as "L_θ old" and inside brackets. The table 5.7 also includes the values of what it would have been the orientation value if only the considered landmark corner were considered (under the column θ_{corner}). This shows the advantage of integrating more estimates. Note that the ground truth for the landmark orientation is 70 degrees.

By looking at the results in table 5.7, the following facts can be outlined:

- The case of landmark L_{11} represents a very favorable condition. This can be identified by a relatively "low" value in all variance estimates, i.e. σ_x, σ_y, and σ_θ, which tells us about the accuracy of the landmark center. In addition, "low" variance values are also associated with, C_2, i.e. the only corner matched. As a consequence, the orientation estimate L_θ is very close to ground truth.
- The case of landmark L_{10} represents a situation more uncertain than the one of landmark L_{11}, (variance values are higher). Consequently, if only one corner is matched the error in landmark orientation is around 2 degrees. Nevertheless, if both the corners are

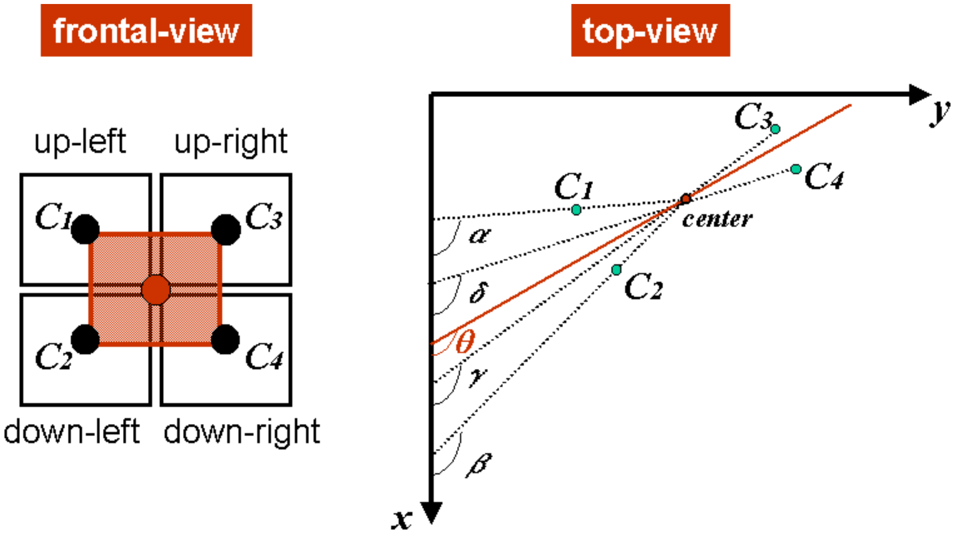


Figure 5.38: The figures represent the proposed technique to estimate landmark orientation, (L_θ), in the case of matching four landmark corner regions. The figure represents the main parameters involved in the equation 5.29 (right-hand), and the landmark and its corner regions surrounded by a red or a black square respectively (left-hand).

considered the orientation estimate is more precise (0.5 degrees error). In this case it is suggested to average the corner estimates since the associated errors are acceptable, (i.e. under a pre-determined threshold).

- The case of landmarks L_6 and L_4 show how a relatively large uncertainty (in σ_y and σ_θ) can be recovered by the proposed weighted average. Note that the result is an error lower than the single estimates.

- The case of landmark L_{14} represents the case when the uncertainty in landmark center is above the pre-determined threshold, (this concerns both σ_y and σ_θ). The consequence is that the landmark is discarded. To discard the landmark seems a reasonable thing to do when testing what would have been the estimated orientation value, i.e. 84.9, representing an error of 14.9 degrees. Note that the estimate related to corner C_4 was anyway not considered in the orientation estimate, falling the computed corner position outside the allowed ”ring volume”, i.e. corner C_4 did not pass the ring-test.

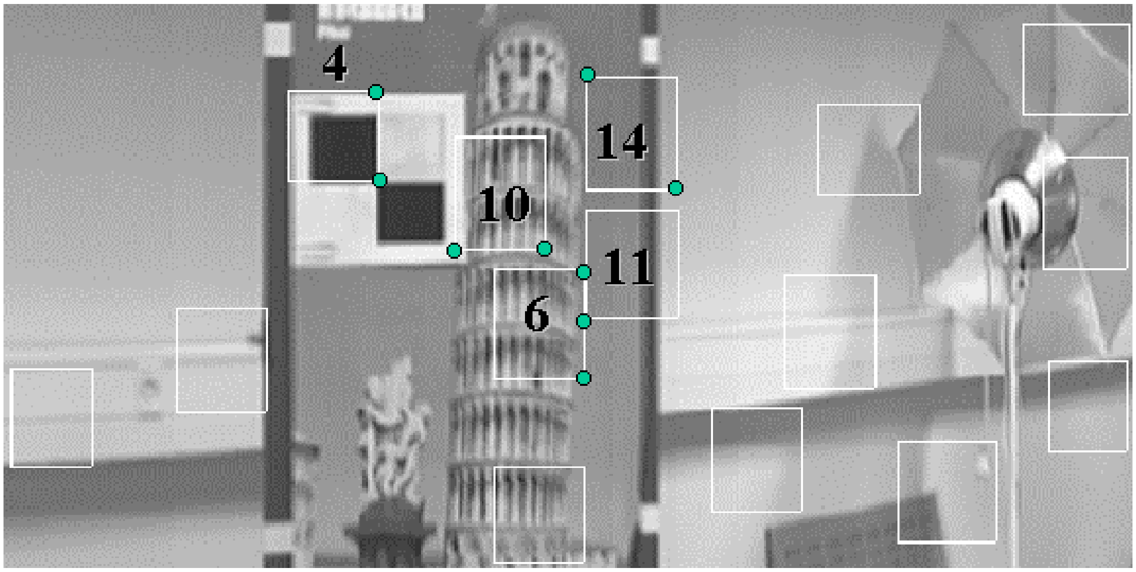


Figure 5.39: The figure shows a region of the new laboratory containing the rotating panel and an "ideal" landmarks. The landmarks which have been extracted in a typical run by the proposed system and that have passed the correspondence test are denoted by a white square with an associate number. The landmark corner which have been matched are denoted by a green circle.

Landmark	reg.	$\mid\sigma_x\mid$ (mm)	$\mid\sigma_y\mid$ (mm)	$\mid\sigma_\theta\mid$ (deg)	θ_{corner}	L_θ	$L_\theta(old)$
L_{11}	$(center)$	72.6	1.6				
	(C_2)	72.7	2.7	0.02	$\beta = 70.1$	**70.1**	(70.1)
L_{10}	$(center)$	72.6	2.9				
	(C_2)	72.6	5.3		$\beta = 71.8$		
	(C_4)	72.6	5.3	0.09	$\delta = 67.1$	**69.5**	(69.5)
L_6	$(center)$	72.7	5.0				
	(C_3)	72.7	2.8		$\gamma = 72.2$		
	(C_4)	72.3	7.3	7.32	$\delta = 63.8$	**70.2**	(68.0)
L_4	$(center)$	74.2	6.2				
	(C_3)	74.7	8.4		$\gamma = 60.8$		
	(C_4)	74.0	3.8	6.4	$\delta = 69.0$	**66.5**	(64.9)
L_{14}	$(center)$	73.1	10.5				
	(C_1)	72.5	7.6	8.10	$\alpha = 84.9$	84.9	(84.9)
	(C_4)	49.6	2.6				

Table 5.7: The table reports the orientation estimates related to landmarks represented in figure 5.39 and matched corners. The accuracy of the orientation estimates can be different depending on many factors (see text).

5.3.2 Landmark Election

The acquired landmark views will be used as reference images during self-localization. This implies that a well defined and low-distorted landmark image will play a much better service than a stretched and distorted one. The same can be said for a landmark which position has been accurately estimated compared to another one which has got a very uncertain estimate. A better reference image as well as a precise positional information will decrease the probability of either mismatches or missed recognition, and the effect of this will be less noise in the estimated robot pose. The proposed landmark election step consequently represents a further attempt to discard candidate landmarks which may represent an unreliable information. In particular, such that they do not satisfy the proposed reliability criteria in terms of *vergence* and *orientation* angle.

Vergence Angle

The main parameters involved in stereo reconstruction are represented in figure 5.40 (for an example room). The larger the baseline of the stereo situation, the more representative is a matched texture patch. However, the larger the baseline the higher the discrepancy between two stereo views, so that it is more difficult to match them. The solution has then to be a compromise between having a large baseline and meeting stereo reconstruction requirements. The consequence of such a compromise is that the distance and the orientation to many potential landmarks could not allow for a robust image acquisition. In particular, the distance and orientation to a landmark has a great influence on how a landmark appears in the camera image-plane.

A parameter to be considered is the *vergence angle*. That is in our case, the angle between the lines passing through landmark center and camera optical center, for two different camera positions. This angle is labeled α_v in figure 5.40. The amplitude of angle α_v depends on the distances between landmark-center and camera-centers and on the distances between the two camera-centers. These distances are in the figure labeled d_1, d_2, and *baseline*, respectively. The angle α_v is important because the larger this angle the more different the landmark views in the "left" and "right" image planes.

An interval of favorable values for the angle α_v, represented by the interval $\left[\alpha_v^{min}, \alpha_v^{max}\right]$. has been experimentally determined in the new laboratory environment in case of the "ideal" landmark. The experimentation was based on introducing in the workspace a certain number of "ideal" patterns all around the workspace and then running the learning procedure presented in section 5.1 and 5.2. The workspace set up is shown in figure 5.41 through a panoramic view, (the same as in figure 5.17 top-row).

The general idea is that for typical indoor environments there would be set a range of "acceptable" values for α_v. Then, it could be possible to verify whether a landmark has got an "acceptable" α_v or not, and then save or discard the landmark accordingly to the result of this test.

The red areas in figure 5.42 illustrate the workspace areas where our landmarks were matched. This set a range of values for the angle α_v considered acceptable. The green areas in figure 5.42 represent then the workspace regions where further landmarks of the same type are imagined to be matched (if extracted). The green areas in fact contain "acceptable" values for α_v. The regions where representative views are instead expected not to be found are

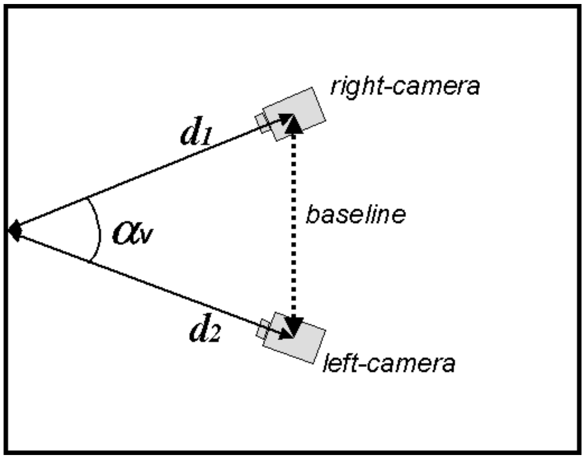


Figure 5.40: The figure shows main parameters involved in a stereo reconstruction. The distance between the landmark center and the camera center is labeled d_1 for the right-camera, d_2 for the left-camera. The distance between the two camera centers is labeled $baseline$. The vergence angle, labeled α_v, depends on d_1, d_2, and $baseline$. The larger the α_v the more different the landmark appearance in the left and right image-planes.

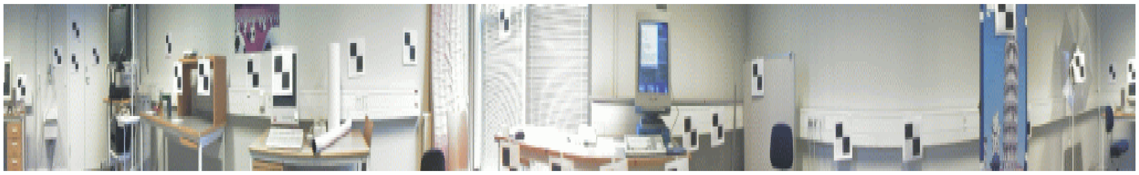


Figure 5.41: The figure shows the experimental setup through a panoramic view. The workspace was filled in with plenty of "ideal" patterns.

those outside the green areas. As an example the angle α_{v1} in figure is considered acceptable whereas the angle α_{v2} is not. The test of vergence angle will then remove landmarks belonging to "unfavorable" workspace regions.

Instead than the vergence angle, the distance between the landmark center and the principal viewing axis[3] could be considered. Still, knowledge on the workspace expected dimensions is required in order to determine a range of "acceptable" distances.

It has to be noticed that the above test does not take in account the orientation of the cameras C_θ^{left} and C_θ^{right}. This is due to the fact that proposed acquisition strategy based on synthesizing panoramic view by cylindrical projections leads to the conclusion that every landmark center can be considered to be contained in the camera optical axis related to its observation.

[3] The principal viewing axis is defined as the axis which is the extension line from one camera optical center to another.

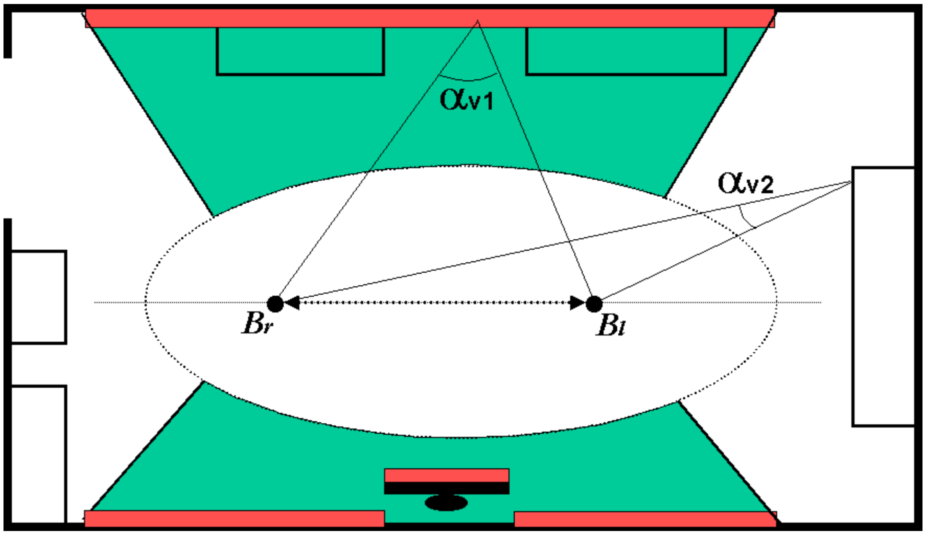


Figure 5.42: The figure shows the workspace floor-map (new laboratory). The red slim rectangular areas in figure (on parts of the walls and rotating panel) illustrates the workspace areas where tested landmark views were matched. This set a range of values for the angle α_v considered acceptable. The green areas in figure represent then the workspace regions where further landmarks of the same type as the tested ones were imagined to be matched. In other words, landmarks belonging to the green areas should be reliably matched. The regions where representative views are expected not to be found are instead those outside the green areas. As an example the angle α_{v1} is considered "acceptable" whereas the angle α_{v2} is not.

Orientation Angle

The landmark orientation (to acquisition baseline) also plays a role in the reliability of the landmark extracted information. It is consequently proposed to test how "favorable" is the angle between landmark surface and baseline for the learned landmarks. This angle in fact tells us about the distortion inherent to the acquisition of landmark appearance in the image-plane.

So as demonstrated by testing observation areas (both in chapter 4 and more accurately in chapter 6), there is a limit in the orientation angle (around 30 degrees) which depends on the landmark texture structure and other things, (see sub-section 6.4.2), above which a landmark is not matched any longer.

It may happen that despite all "consistency checks" that a landmark has been through, its estimated orientation angle is not consistent or otherwise acceptable, for a reliable use of the landmark during the self-localization phase. This happens when the estimated landmark orientation angle is above the estimated threshold. This in fact indicates that "something is wrong". Thus, the landmark will be discarded. Note that there can be cases of landmarks having acceptable vergence angle but an unacceptable orientation angle.

5.4 Learning Navigation Strategy

The proposed solution for automatic learning of an environment model based on visual landmarks has been presented in the previous sections: Landmark View Acquisition (5.1), Landmark Pose Computation (5.2), and Model Refinement (5.3). This section is to briefly discuss "how" and "when" the proposed learning method can be applied. In other words, which is the *learning navigation strategy* for the proposed approach to autonomous robot navigation.

The proposed learning method requires:

- the acquisition of at least two environment views (panoramas) from different camera positions, such that a part of the workspace has to be visible from the chosen positions.
- the knowledge of camera 3D position at the moment the views were acquired.
- the camera focal length in pixels.

The robot initial position is not necessary, unless there is a need for relating the robot position to an external map or other kind of information which could be provided.

Since the robot is supposed to move to learn the environment, it becomes very important to investigate which is a convenient navigation strategy. The could be special procedures to be proposed for driving the robot during the learning phase even if the environment is unknown, however, this topic is not dealt with in this thesis, and we just consider in this section general learning paths which may "naturally" arise during navigation.

In general, the most important aspect is that the navigation strategy should allow the system to learn an environment model, and at the same time, not accumulate high uncertainty. The latter is very important since the error on computed landmark poses depends on the error in robot position. Initially, when no knowledge is yet available about the environment, it is expected that the robot will estimate its traveled distance using internal sensors, e.g. the odometry. The use of this type of sensory input consequently means that in order to maintain a low positional uncertainty the traveled distance should be a short range (less than 5 meters) and turns should be avoided.

The robot system is expected to have learned at least a few landmarks after the first two acquisitions. This is believed possible for most of the possible learning trajectories since planar landmarks are abundant in typical indoor environments, and since the system is relying on 360 degrees panoramic views. Once a few landmarks have been learned, the system can make use of them when moving towards the next acquisition position, that is, the system could use the learned information for self-localization while traveling to the next acquisition position.

As a consequence of having a few camera locations for the acquisition, and to meet stereo reconstruction requirements, the distance and the orientation to many potential landmarks could not be convenient for a robust image acquisition. Nevertheless, there should always be landmarks which are suitably located for being landmarks, and, as described in chapter 6, the proposed algorithm is be able to recognize this situation.

In cases when the current acquisition configuration does not provide sufficient information, the proposed learning procedure can again run at the next robot position. In general, depending on the current knowledge there could be a need for running the learning procedure on multiple acquisition positions.

The above paragraphs have in short described the proposed navigation concept. That is an approach to autonomous robot navigation which relies on two phases: Learning and Self-Localization; which may dynamically succeed each other, depending on the current needs. The diagram in figure 5.43 illustrates the proposed learning navigation strategy, which primary computational steps being described by the following algorithm.

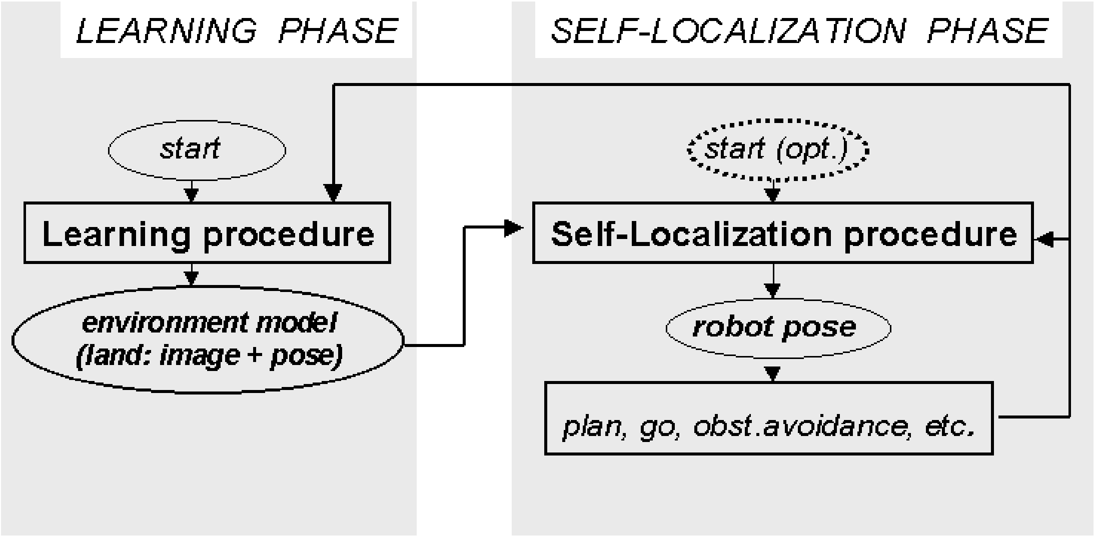


Figure 5.43: The figure illustrates the proposed learning navigation strategy: Learning-phase, i.e. automatic acquisition of visual landmarks; and Self-Localization phase, i.e. estimation of robot pose during navigation.

```
1. $i = 0$ ; $R_i = 0$ or $R_i$ ="init-pos" ; DB ="empty" ; acquire panorama
2. $i = i + 1$ ; "move to $R_i$"[4] ; acquire panorama
3. "run LEARNING", baseline $= (R_i, R_j)$, j=i-1,..,0 ; DB = "all learned landmarks"
4. "Is DB sufficient for accurate navigation towards the goal? "
    5a. "NOT" ; go to step 2.
    5b. "YES" ; Repeat: "run SELF-LOCALIZATION" during navigation towards goal
                  Until: "positional uncertainty is "low" & DB covers the navigating area
6. $R_i$ ="current-position" ; go to step 3.
```

[4] R_i should be selected such that is not far away and the position could be reached with low-uncertainty.

How many panoramic-views should be synthesized? and from which positions? This is a topic to be thoroughly investigated in future research. In the experimentation runs presented in this thesis, the robot positions for the acquisition were simply chosen in a way that most of the environment was visible. This type of situation could naturally arise during a learning phase, (or could be made to happen with a little human help). The figure 5.44 shows the learning paths in the presented experimentation.

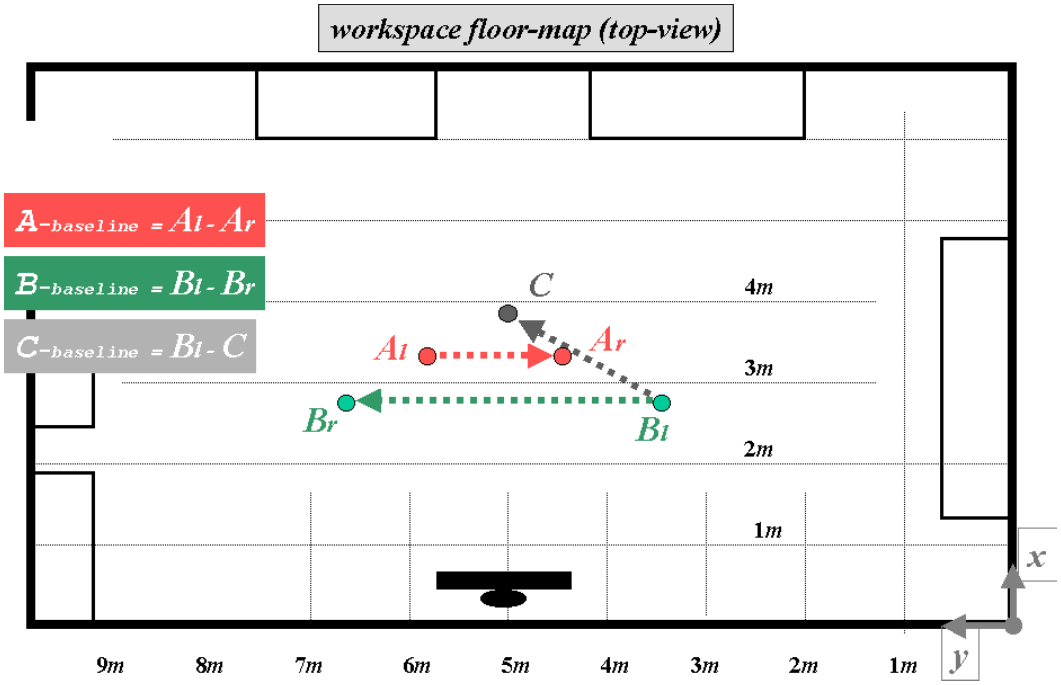


Figure 5.44: The figure shows the learning paths (the baselines) for the learning experimentation sessions presented in this thesis.

The figure 5.45 illustrates possible ways the robot could proceed towards its goal, i.e. the room door, from two of the baselines used for the experimentations, (represented by the A- and the C- *baselines*). In particular, the system could either consider sufficient the learned information to reliably reach the door (top-row), or opt for an additional ”learning stop” at position D and E respectively (bottom-row). Note that in the latter, the second learning baseline is represented by $A_r - D$ in case of ”path A and $C - E$ in case of ”path B”, (both the two baselines are represented in by green arrows in figure 5.45 bottom-row). The learning baselines $A_l - D$ and $B_l - E$ could also be considered to learn additional landmarks in case the distance is not considered ”prohibitive”.

The figure 5.46 illustrates more examples concerning possible learning navigation strategy which are not related to presented experimentations.

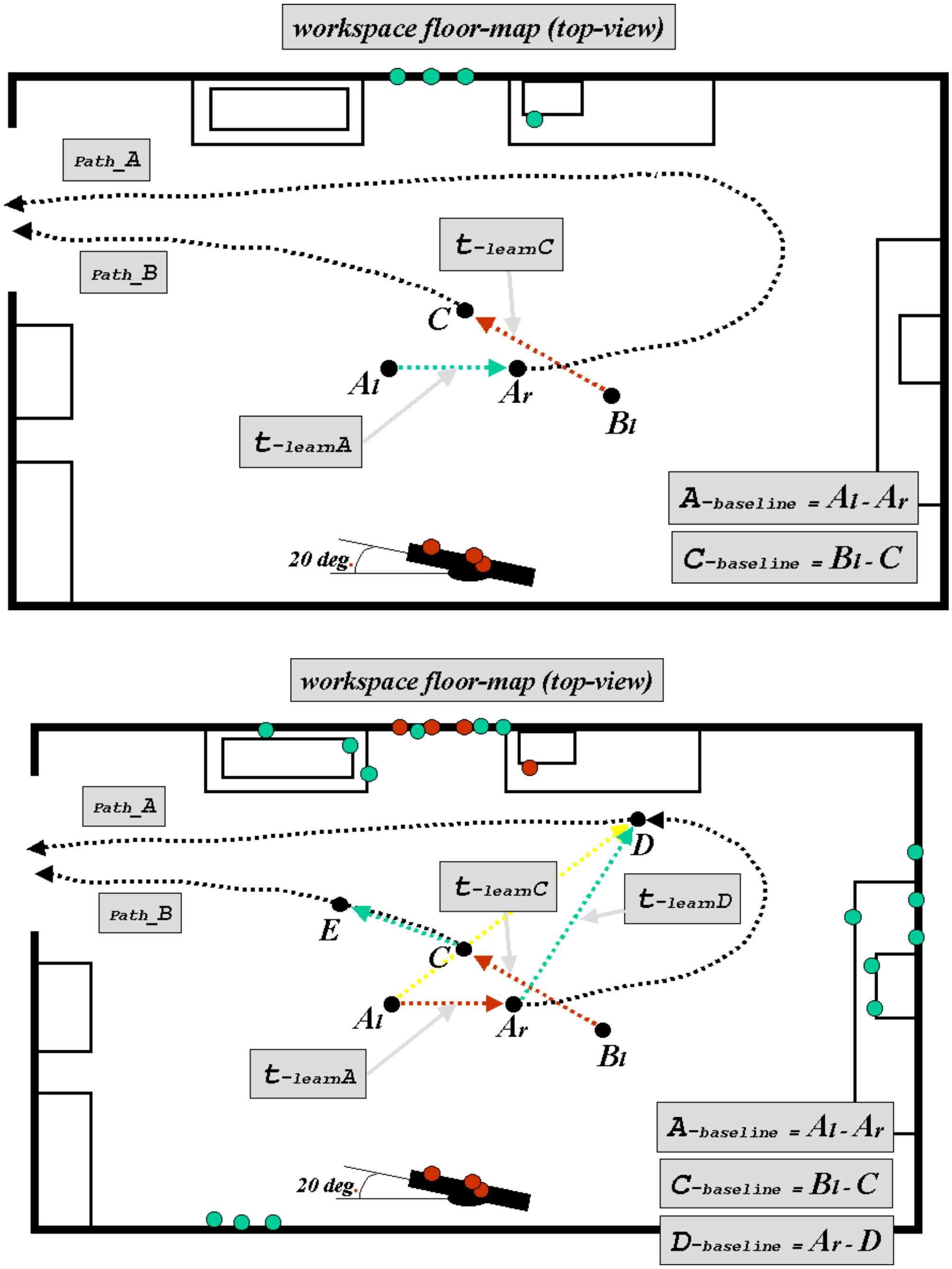


Figure 5.45: The figures show the new-laboratory workspace floor-map. The figures illustrates possible ways the robot could proceed towards its goal, i.e. the room door, from two of the baselines used for the experimentations, (represented by the *A*- and the *C*- *baselines*). In particular, the system could either consider sufficient the learned information to reliably reach the door (top-row), or opt for an additional "learning stop" at position D and E respectively (bottom-row). The green and red circles represent the learned landmarks related to the green and red baseline respectively.

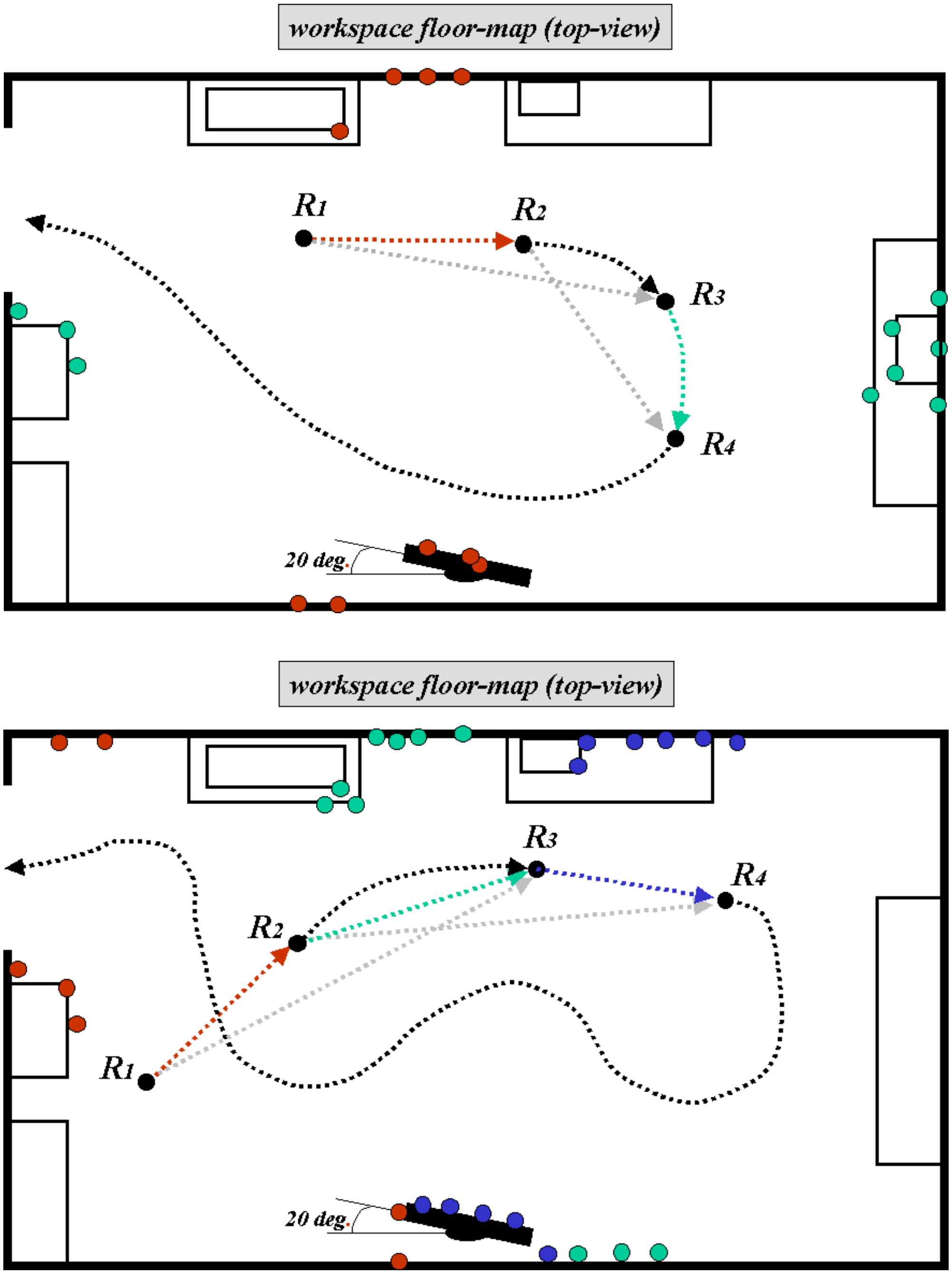

workspace floor-map (top-view)
R1
R2
R3
R4
20 deg.
workspace floor-map (top-view)
R3
R4
R2
R1
20 deg.

Figure 5.46: The figures show the new-laboratory workspace floor-map. In particular, the figures illustrates examples concerning possible learning navigation strategy which are not related to presented experimentations. The R_i represent robot positions. The green and the red arrows the learning baselines. The green and red circles consequently represent the learned landmarks related to the green and red baseline respectively. The grey arrows represent possible additional baselines which could be considered.

The proposed learning navigation strategy provides a robot with an high level of autonomy which makes the system capable of accomplishing navigation task in unknown or changing environments. What the proposed learning requires is that there is a sufficient amount of information, (environment features), which remain stable through the learning and navigation time. The basis concept in fact is that the system relies, during navigation, on what it has previously learned. Nevertheless, the most interesting, useful and very potential characteristic of the proposed system is that it allows for changes in the environment, and this thanks to the proposed learning navigation strategy.

An application example for the proposed system is a warehouse or a goods deposit, where a fork-lift robot, (e.g. as the one represented in figures 5.19 and 5.23), could be employed to fetch and carry boxes and packages, for example from the track which just arrived to a specifics warehouse sectors. The warehouse environment would then continuously change its appearance due to the new boxes which have come in (transported by the same robot). Though, the system would be able to cope with this variations by learning new landmarks in the area which have changed, while still relying on those landmarks, previously learned, related to the warehouse areas which have not changed. This represents the way the system will progressively adapt its knowledge to environment modifications.

5.5 Summary

The learning step was believed a fundamental step in order to make a mobile robot fully autonomous. Consequently, a learning phase has been proposed to our system, and this chapter has described the proposed approach.

The information the system is required to learn for robot self-localization consists of landmark visual appearance and positional information, (representing the proposed environment model).

The proposed method for automatically learning the environment model is characterized by a sequence of computational steps which are proposed in order to compute the information required for robot localization while testing accuracy and reliability of the learned information.

In particular, a large amount of visual information is acquired at the beginning, and then filtered and validated through a sequence of "validity checks" aimed to reveal the presence of an unreliable or "weak" information, to then only leave the "elected" landmarks, i.e. the most reliable.

The main computational steps for the learning phase are:

- **Landmark View Acquisition** (section 5.1)

  This step consist of:

  1. acquire different views of the environment surrounding the robot by panning the on-board camera over 360 degrees while the robot is standing at a known position;
  2. synthesize environment panoramic-views from the acquired images in one view by using cylindrical projections;
  3. extract salient panoramic regions (high-contrasted and unambiguous) by means of a proposed attentive scheme. The extracted regions represent candidate landmarks through their visual appearance.

- **Landmark Pose Computation** (section 5.2)

  This step consists of:

  4. test correspondences between candidate landmarks by matching their appearance in different panoramic-views;
  5. estimate landmark 3D position by stereo-triangulation;
  6. estimate orientation of the surface containing the landmark by a proposed technique based on multiple positional information related to landmark corners.

- **Model Refinement** (section 5.3)

  This step consists of:

  7. estimate landmark positional uncertainty by the analysis of the error propagation in order to improve landmark orientation estimate, discard very uncertain landmarks, and classify maintained landmarks based on the estimated covariance. The latter to represent a useful input for landmark triplet selection during navigation;
  8. test consistency and limit conditions for the learned positional information associated to landmarks through the analysis of some critical angles (vergence and orientation angle);
  9. store in a system database the landmark which have been elected to be used by the system in the self-localization phase.

A brief discussion about "how" and "when" the proposed learning method could be applied for autonomous robot navigation, together with a general algorithm and some illustrations exemplifying the general application context of the proposed learning and navigation strategy, was then reported in section 5.4, **Learning Navigation Strategy**.

Chapter 6

Automatic Recognition of Self-Learned Landmarks

Full autonomous navigation requires a robotic system to be self-reliant. In case of the proposed navigation strategy (based on vision and natural landmarks) this means an automatic acquisition of visual landmarks during the learning phase and an automatic recognition of learned landmarks during the localization phase. The proposed method for automatic recognition of landmarks, (presented in section 4.1), is consequently required to perform also in case landmarks have been self-learned by the system.

This chapter proposes an approach for automatic recognition of self-learned landmarks. In particular, in case of landmarks which have been learned by the method proposed in chapter 5. The process of automatic learning of landmarks introduces additional challenges to automatic recognition due to the adopted learning method and to the uncertainty associated to learned information. It is consequently important to design the automatic recognition method according to the acquisition.

The previously proposed automatic recognition method (see section 4.1) is then "replaced" by a new method which better exploits the possibilities of the proposed learned method, and better copes with the challenges of an automatic learning. The main "guidelines" of the recognition approach described in section 4.1 are however kept: landmark visual prediction is obtained by a form of re-projection based on previously acquired references and positional information; and the landmark match is still based on template matching (normalized cross-correlation).

Nevertheless, a new landmark "re-projection" method is proposed which synthesizes realistic landmark virtual views and at the same time exploits different reference views for the same landmark. This allows the system to better exploit proposed learning strategy (chapter 5) and to better deal with robot and landmark positional uncertainties.

6.1 Core Idea and Argumentation

The problem of automatic recognition of landmarks is often studied in the literature related to the acquisition strategy. For example, landmark recognition is often proposed at appearance level so that both the spatial representations (for the recognition and the acquisition) are close to each other, (and to the sensor). The advantage is in this case that the representations are grounded on the perceptual ability of the robotic system. For example, the landmark recognition process often "directly re-uses" acquired landmark-views, as template for the visual prediction, (Zingaretti and Carbonaro [153], Balkenius [9]). It is consequently believed important to our system that the design of the landmark recognition method should be according to the method proposed for the automatic learning of landmarks, (see chapter 5).

In case of a "direct re-use" of acquired landmarks views, the viewpoint invariance of the landmark plays an important role. Thus, the extension of the "observation area" becomes an important parameter which determines the required number of landmarks in the environment. Note that a "direct re-use" of landmark acquired views as template for the landmark matching leads to observation areas smaller than those related to proposed *"landmark reprojection"* concept, (section 4.1), though, special templates can be used to enlarge observation areas, e.g. Balkenius' elastic template [9]. It is consequently believed advantageous to propose the concept of "landmark reprojection" combined to the acquisition strategy proposed in chapter 5.

The automatic recognition method proposed in section 4.1, represented the optimal solution for acquisition of reference-views related to planar objects in case of acquisition performed by manual training. In fact, the possibility of capturing occlusion-free frontal-views of landmarks, i.e. the most representative view for a planar object, led to the fact that these views could be used as reference views for the reprojection. In case of an automatic learning performed while the robot is navigating the environment, it is instead unlikely to capture an ideal observation condition. Rather it acquires several non-ideal landmark views from different viewpoints. In chapter 5 this problem has been discussed and a solution proposed in order to select the best views among those acquired, or in better terms, to extract best positioned landmarks according to acquired views. The problem we want to investigate in this chapter is then more precisely, how to best exploit views acquired by the method proposed in chapter 5, in order to achieve reliable and automatic landmark recognition.

The innovative idea this thesis proposes for automatic recognition of self-learned natural landmarks is the transfer of some methodologies proposed for realistic visualization of virtual views to reliable recognition of observed objects. We could also say in short that the goal of a realistic visualization has turned towards a reliable match.

The proposed idea came natural to the author of this thesis when dealing with the problem of an accurate (and thus realistic!) landmark *visual prediction*. In particular, the attention was focused on the possibility offered by methods proposed in the realistic visualization literature, of an advantageous exploitation of image correspondences for the transfer of pixel values to a new view point, (virtual).

A collection of reference images which captures the appearance of a landmark-object from different viewpoints may for example allow for higher fidelity in visual predictions, if textures could be estimated view-dependently. This technique is in fact expected to better reproduce local illumination effects and object visible aspects.

In addition, the general approach followed in realistic visualization literature (image-based rendering in particular) of first acquiring sample-images, and later using them to generate virtual views, sounds very appropriate and convenient to the proposed navigation approach, which consists of first learning the environment and later using learned information for the purpose of recognition (and self-localization).

Such an approach would definitely tie together the recognition process (generated visual predictions) and the automatic acquisition process, keeping in addition, the advantage of the proposed *"landmark re-projection"* technique (which in turns means "large" observation areas), and no need, in principle, for a full geometric reconstruction.

The main issue then becomes: what is the reliability of a match involving real and virtual views?

The answer need to be searched in: (1) the method used for synthesizing visual predictions; (2) the way the system deals with uncertainties in positional information related to landmarks and robot.

In summary, the main aspects supporting the proposed idea of transferring some of the methodologies of the realistic visualization to improve image-recognition performance, in a system for autonomous robot navigation, are:

- The exploitation of a collection of reference images which captures different appearances of a landmark from different viewpoints allows for higher fidelity in landmark visual predictions, (multiple reference views may encompass areas not entirely observed in only one view).
- The general approach followed in the realistic visualization literature of first acquiring sample-images, and later using them to generate virtual views, seems to be very appropriate and convenient for the proposed two-phases (learning and localization) robot navigation approach.

The application of methods from realistic image synthesis to robot navigation based on naturally occurring visual landmarks observations, potentially has a tremendous impact in terms of recognition flexibility and match reliability. In particular, in case when landmarks represent objects having arbitrary shapes. A landmark shape not restricted to planar represents a natural future step of the research in the topic proposed in this thesis.

Nevertheless, also in case of planar or "almost-planar" landmarks, as the ones utilized in this thesis, there is a clear advantage in applying techniques from realistic visualization of virtual views. For example figure 6.1 shows the case of a landmark representing part of a computer monitor. In this case, landmark appearance changes a lot when the landmark is observed from different viewpoints, so that the application of methods from realistic virtual view synthesis results in a higher fidelity of synthesized landmark appearance, compared to what is achieved by the projection of a single reference-view. This allows for a higher recognition performance and a wider observation area, (see how the observation area Obs_1 in figure, is imagined to be extended to also include area Obs_2).

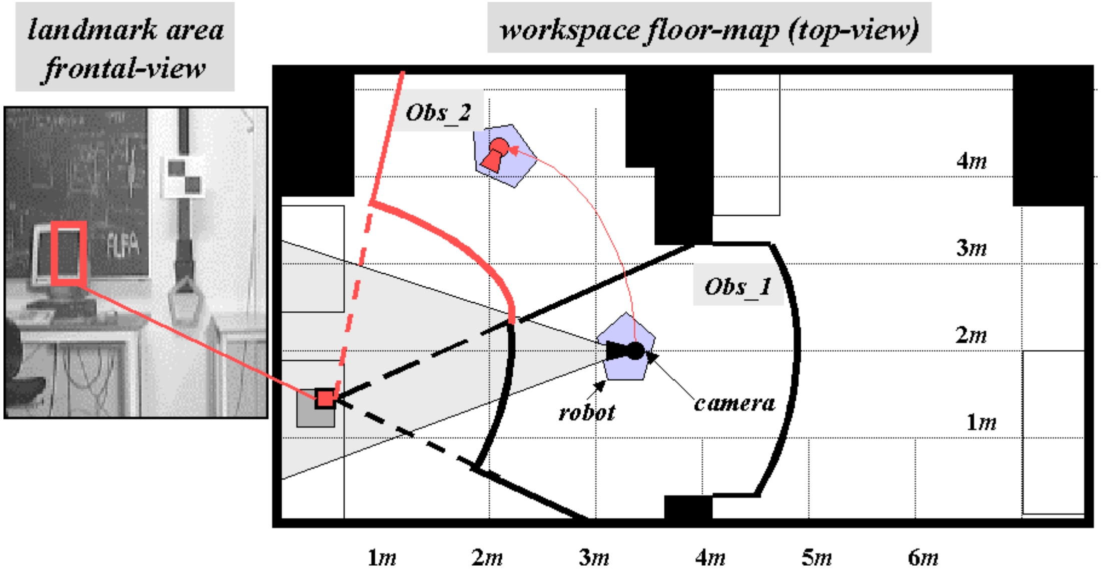


Figure 6.1: The figure shows a conceptual example of how observation area may vary (from Obs_1 to $Obs_1 \cup Obs_2$) when landmark visual prediction is generated by a realistic image synthesis based on multiple reference views (from different viewpoints).

6.2 Realistic Visualization of Virtual Views

Realistic Visualization of Virtual Views is a new field of research which has received increasing attention in recent years. It is strictly related to the increased popularity of virtual reality and the spread of its applications, hence, to the demand of increasing realism into graphically generated sceneries and to ease the modeling process.

Computer Graphics allows us today to visualize in real-time innumerable and amazing scenery with no limits on viewpoint and viewing direction. Despite of this, the artificial nature of data makes the visualized virtual sceneries not realistic enough. Not-realistic in the sense that a synthetic scene is easy to discriminate visually from a natural scene. In addition, all the physical phenomena involved in a realistic scene visualization can be difficult or impossible to re-create. A more realistic synthesis of virtual sceneries is consequently highly desired and it finds applications in all the usual domains of image synthesis, hence, all sectors which take benefits from virtual reality.

There are many factors which contribute to a more realistic view synthesis, among them geometric modeling, object textures, and illumination simulation. Since vision is the dominant human sensory modality, a high-fidelity visual feedback is considered a primary aspect in order to involve human beings in a way that make them feel and act as if they were in their natural environment, (thus, to achieve the effect of a visual presence).

The main goal of realistic visualization of virtual views is a high fidelity representation of visualized sceneries. In particular, to provide realism by the transfer to the new view, of all the physical phenomena captured in the reference photographs, (i.e. the transfer of photo-realism). It is in fact the summation of all the natural phenomena represented in a generated view, which provides the user with a sense of presence (based on the visual realism).

Currently, there are two dominant approaches to the generation of virtual views:

- Model-Based Rendering
- Image-Based Rendering

These two classes of approaches are in the following briefly introduced and commented (subsections 6.2.1, 6.2.2, 6.2.3, respectively). A brief review of selected approaches in the field can be found in Appendix A.

6.2.1 Model-Based Rendering

Model-based rendering represents the traditional approach for generating virtual views of an object or a scene. It fully relies on a geometrical description of objects or scenes, (e.g produced by CAD modelers or from real data acquisition), from which images can be rendered from any desired viewpoint. Realism can be enhanced by applying texture maps (images from real or synthetic scenes), shading algorithm, etc., on surfaces of the geometric 3D models. Main processing steps of model-based rendering are described in figure 6.2.

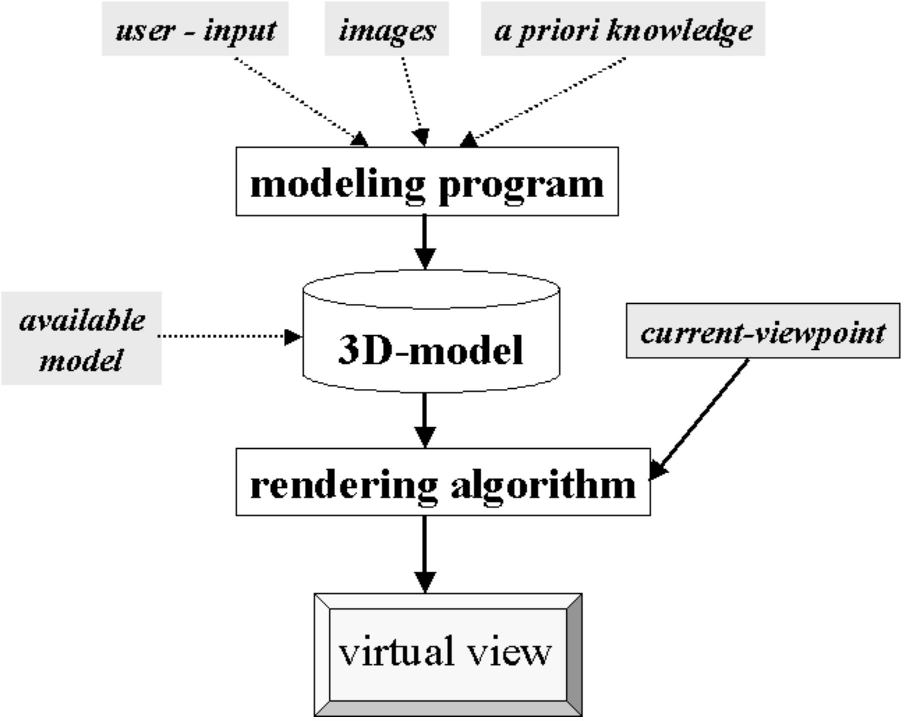


Figure 6.2: Model-based rendering: typical computational steps. Dashed arrows symbolize optional inputs.

The advantage of a model-based approach is a compact data representation which in principle consists of accurate models and few images, and it has the flexibility to generate any new view which may be required. In addition, specialized graphic hardware is commercially available and can be exploited.

The main drawback of model-based approaches is the reliance on the availability of the models. The geometric model-reconstruction may represent a problem in case of complex shaped objects, as well as for the texture reconstruction. In addition, the computational complexity of geometries and illumination simulation, make the rendering strongly depending on model complexity and hardware capability. Finally, to accurately reconstruct the real-world it is a tedious or even impossible task because of the complexity of the occurring phenomena, which makes high-fidelity visualization of real-scene hard to achieve with a model-based rendering and often leaves the user with an impression which is far from a satisfactory realism. Figure

6.4 summarizes the main characteristics of model-based rendering approaches.

6.2.2 Image-Based Rendering

Image-based rendering represents an emerging and competing means of creating virtual views. In contrast with model-based rendering, image-based approaches primarily rely on a set of pre-acquired real-images, and they utilize special techniques to interpolate the acquired images, or reproject textures from source images to target image, in order to produce *novel* virtual views. This powerful new class of approaches generates photo-realistic virtual views, and compelling transitions between reference images can be generated without an explicit geometric representation. The main processing steps of image-based rendering are described in figure 6.3. An imagined example is shown in figure 6.5.

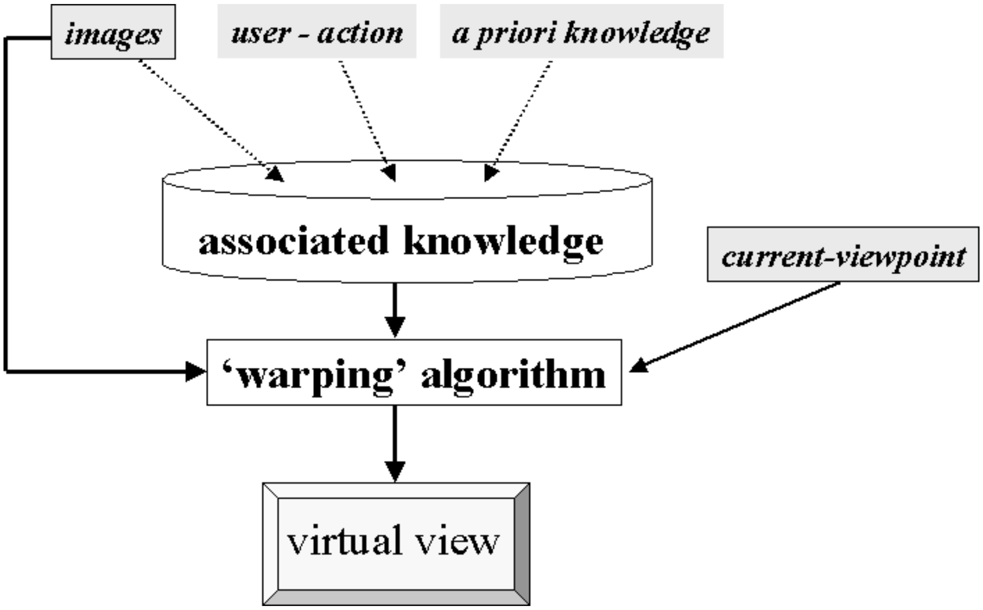
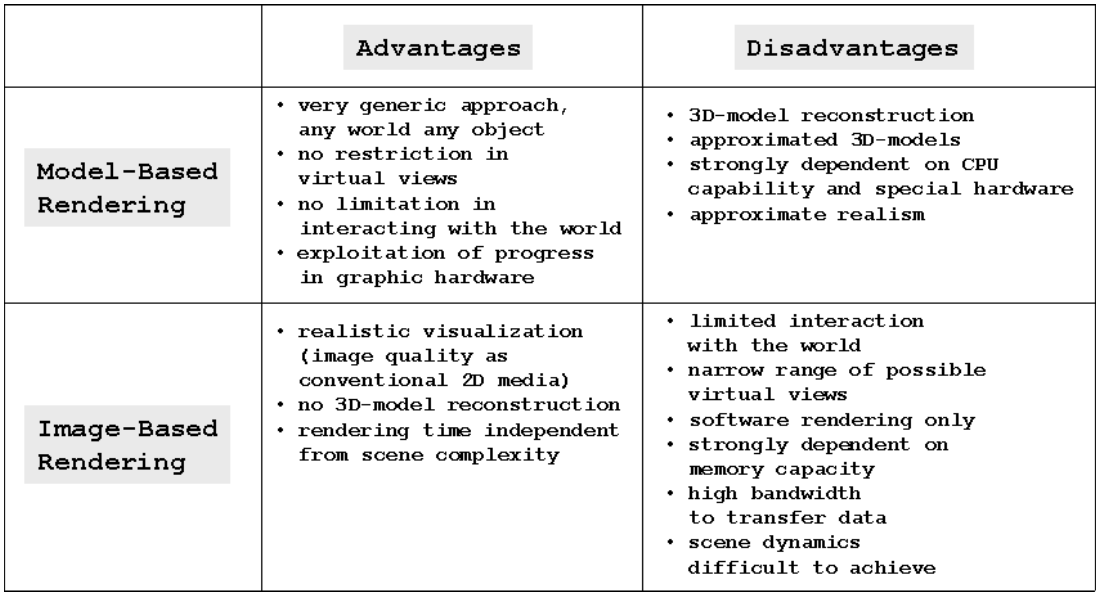


Figure 6.3: Image-based rendering: typical computational-steps. Dashed arrows symbolize optional inputs. The "associated knowledge" may consist of correspondences information, depths, epipolar relationships, etc.

The main advantages of image-based approaches are: the realistic nature of the resulting images, a rendering time independent of scene complexity, and no need for geometric model reconstruction. The quality of created novel views is based on the quality of the acquired images and not on the complexity of the scene and object represented.

The major limitations of image-based approaches are: narrow range of supported virtual-views (i.e. it is often required that reference views lie close to each other), and consequently a need for a relatively large number of images and memory capacity, disadvantages in networking due to the bandwidth required to share an image-based virtual-world, and unavailable specialized hardware. Figure 6.4 summarizes main characteristics of image based rendering techniques.



	Advantages	Disadvantages
Model-Based Rendering	• very generic approach, any world any object • no restriction in virtual views • no limitation in interacting with the world • exploitation of progress in graphic hardware	• 3D-model reconstruction • approximated 3D-models • strongly dependent on CPU capability and special hardware • approximate realism
Image-Based Rendering	• realistic visualization (image quality as conventional 2D media) • no 3D-model reconstruction • rendering time independent from scene complexity	• limited interaction with the world • narrow range of possible virtual views • software rendering only • strongly dependent on memory capacity • high bandwidth to transfer data • scene dynamics difficult to achieve

Figure 6.4: The table summarizes advantages and disadvantages of image- and model- based rendering techniques.

6.2.3 Summary and Comments

A number of techniques have been proposed for image- and model- based rendering. Image-based rendering is usually applied to static environments (e.g. Chen and Williams [28], Seitz and Dyer [126]), whereas model-based rendering is often proposed for dynamic scene visualization, (e.g. Moezzi et al. [106], Kanade et al. [77]). However, authors have also proposed both the two approaches for the same application context, (e.g. Blanc, Livatino and Mohr [14], [16]).

A number of techniques have also been proposed for approaches which lie in between the two concepts. For example, Debevec, Taylor and Malik [43] show how the number of pre-acquired images required by an image-based approach can be reduced to a sparse set if approximated geometric models of represented objects are provided.

There have been different ways of classifying image-based approaches and distinguishing between those approaches and the model-based ones. In fact, there is not an unanimous consensus about how to categorize image-based approaches, nor a clear edge between image- and model- based rendering techniques. For example, Buehler et al. [18] generalizes many current image-based rendering algorithms as ranging from Light-Field Rendering (Levoy and Hanrahan [85]), i.e. many images with no geometric information, to View-Dependent Texture-Mapping (Debevec, Taylor and Malik [43]), i.e. few images with a geometric model.

In general, it is possible to summarize that previous work in the field of image-based rendering reveals a "continuum" of representations, (which might include model-reconstruction and model-based rendering), based on the tradeoff of many aspects. Among them: number of required input images, motion assumed for the virtual-camera, knowledge about the scene geometry, depths and correspondences, the way pixel are transferred etc. Previous work on the field is usually classified by the authors depending on which is the aspect they would like to focus on in their contribution.

In their classification of image-based rendering H. Shum et al., [130], propose three categories according to how much geometric information is used: no-geometry, implicit geometry (i.e. correspondences), and explicit geometry. D. Forsyth and J. Ponce, [54], also propose three categories but based on the type of approach: volumetric reconstruction, points transfer, and light-fields. L. McMillan, [102], proposes to distinguish approaches based on the way images have supplemented the image generation process: images to represent approximations of scene geometry, images in a database to represent different environment locations, and images as reference scene models from which to synthesize new views. S. Kang [78] proposes a categorization primarily based on the nature of the scheme for pixel indexing or transfer: non-physically based image mapping, mosaicking, interpolation from dense samples, and geometrically-valid pixel reprojection.

It is on the last class of techniques in the S. Kang categorization, (geometrically-valid pixel reprojection), we have focused our attention. In particular, the interpolation of cylindrical panoramic images is considered very appropriate and convenient for autonomous robot navigation, and it allows for a geometrically-valid pixel reprojection. The latter is an important characteristic since the aim of proposed use of realistic view synthesis is matching real observations, rather than pure visualization.

The next section will thoroughly describe the proposed method, while (as previously mentioned) a brief review of selected approaches in realistic visualization of virtual views, involving both image- and model- based renderings, together with a comparison among them, can be found in appendix A.

6.3 Realistic Visual Prediction of Landmarks

The summation of natural phenomena which can be synthesized in realistically generated virtual views may provide the user with a strong sense of presence (based on the visual realism). This can be the case, even if not all the original physical effects are "correctly" transferred to the newly generated view. A physically-based image mapping is consequently not always necessary in order to provide visual realism. However, when the goal of virtual-view synthesis is not a realistic visualization but, as in our case, the match of real observations, a physically based image mapping represents an important factor to take into account, and it has a great influence on the way the image-matching process should be designed, (e.g. if it should be feature- or correlation- based).

The class of image-based rendering techniques characterized as *"geometric-valid pixel reconstruction"*, [78], typically uses relatively small number of images because of the application of geometric constraints, (either recovered at some stage or known a priori), to reproject image pixels appropriately at a given camera viewpoint. The geometric constraints can be of the form of known depths or correspondence values, epipolar constraints between pairs of images, or trilinear tensors that link correspondences between triplets of images. The geometric constraints can also be exploited to solve the visibility problem, (i.e. when an object or scene surface appears in front of another object or surface, even if it should lie behind).

In the literature different possibilities can be found within geometrically-valid pixel reprojection, depending on the chosen method and available information. These are mainly characterized by the use of trilinear tensors, (Shashua [127], Avidan et al. [8], Hartley [66]) and fundamental matrix, (Faugeras [51], Leveau et al. [84]). The image reprojection is very often based on a direct exploitation of known depths, correspondences, (Chen and Williams [28], Chang et al. [26], Seitz and Dyer [126]), epipolar constraints, (McMillan and Bishop [103], Kang and Szeliski [79]), etc.

It is in the exploitation of epipolar constraints, that we have focused our attention and got inspired concerning the transfer of pixel values between cylindrical reference-views to a new viewing position. In particular, based on recent developments in image-based rendering involving the use of the Plenoptic function, (which describes light rays visible at any point in space), and on cylindrical panoramic images, (McMillan and Bishop [103], Kang [78]), it is proposed the *interpolation of cylindrical panoramic images.* The exploitation of cylindrical panoramic views for the purpose of mobile robot navigation is receiving increasing attention in the recent years, (Yuen and MacDonald [150]). In addition, some authors have very recently announced as their future research activity in mobile robotics, the synthesis of virtual views based on cylindrical panoramic references (Bunschoten and Kröse [19]).

Cylindrical panoramic images can "naturally" be acquired by a robotic system during its navigation in a learning phase, and they can represent the basis from where landmark reference-views can be extracted. This has already been demonstrated in chapter 5. Then, during the self-localization phase, the proposed interpolation of cylindrical reference-views can be used to render realistic visual predictions and to match current observations. Figure 6.6 visually describes the transfer of texture values from cylindrical reference-views to the visual prediction.

The transfer of pixels to the new (virtual) view by interpolation of cylindrical landmark reference-views is proposed to be performed in the following way.

1. *Cylindircal Pixel Transfer.* I.e. mapping geometrical pixel correspondences between reference landmark views and visual prediction. For every pixel in the visual prediction, the corresponding pixels in the reference views are calculated by estimating the angular disparity. This mapping follows the idea of "plenoptic transfer", (McMillan and Bishop [103]).

2. *View-Adapted Texture Mapping.* I.e. mapping texture correspondences between reference views and visual prediction. For every pixel in the visual prediction, its texture value (light intensity) is calculated by a weighted average of corresponding texture values in the reference views. This follows the idea of view-dependent texture mapping, (Debevec et al. [43]), where texture in the visual prediction can primarily be based on reference-images taken from viewpoints which are the closer to current observation. Closer images are in fact expected to best approximate landmark visible aspects and local illumination effects present in current camera observation.

It is based on the realism of virtual-views generated by the proposed method that we aim to improve the match between landmark visual prediction and current observation. The proposed technique thus represents the system's "answer" to the issue of synthesizing reliable visual predictions.

6.3.1 Cylindrical Pixel Transfer

The goal of cylindrical pixel transfer is an automatic and reliable estimate of pixel correspondences between reference-views and visual prediction. The available knowledge consists of camera focal length in pixels, landmark position, orientation and reference-views, and an estimate of current robot pose. In addition, there is the knowledge that learned landmarks lie on planar or "almost planar" surfaces.

The basic concept for interpolating cylindrical panoramic images is shown in figure 6.6. The figure visually describes the transfer of texture values from cylindrical reference-views to the visual prediction for an example landmark, (the "ideal" landmark). The top-right image represents the *"current observation"* where the *visual prediction* represent the view re-mapped from reference cylindrical images according to the current estimate of camera viewpoint. The visual prediction is then used as template in an *image correlation* process to locate the landmark precisely in the image-plane

The basic concept for interpolating cylindrical panoramic images is equivalent to computing 3D points from image correspondences and projecting them to a new target image. McMillan and Bishop, [103], devised an efficient method for transferring known image disparity values between cylindrical panoramic images to a new virtual view. Their approach uses the *angular disparity* (related to each cylindrical pair) to automatically generate warps that map reference views to arbitrary cylindrical or planar views.

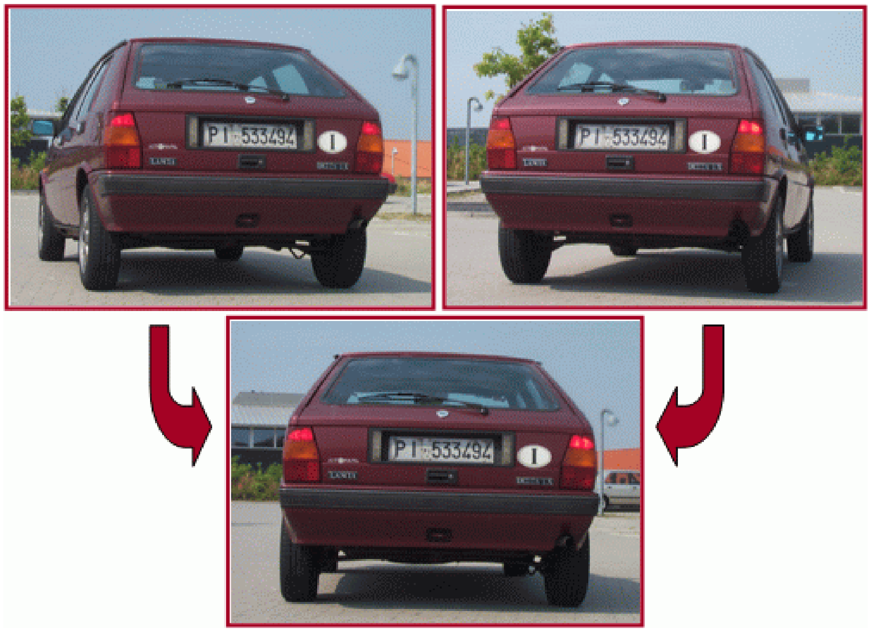


Figure 6.5: The process of image-based rendering through an example: the figure bottom-row represents the expected result: a novel virtual view ("visual prediction") synthesized from two reference views.

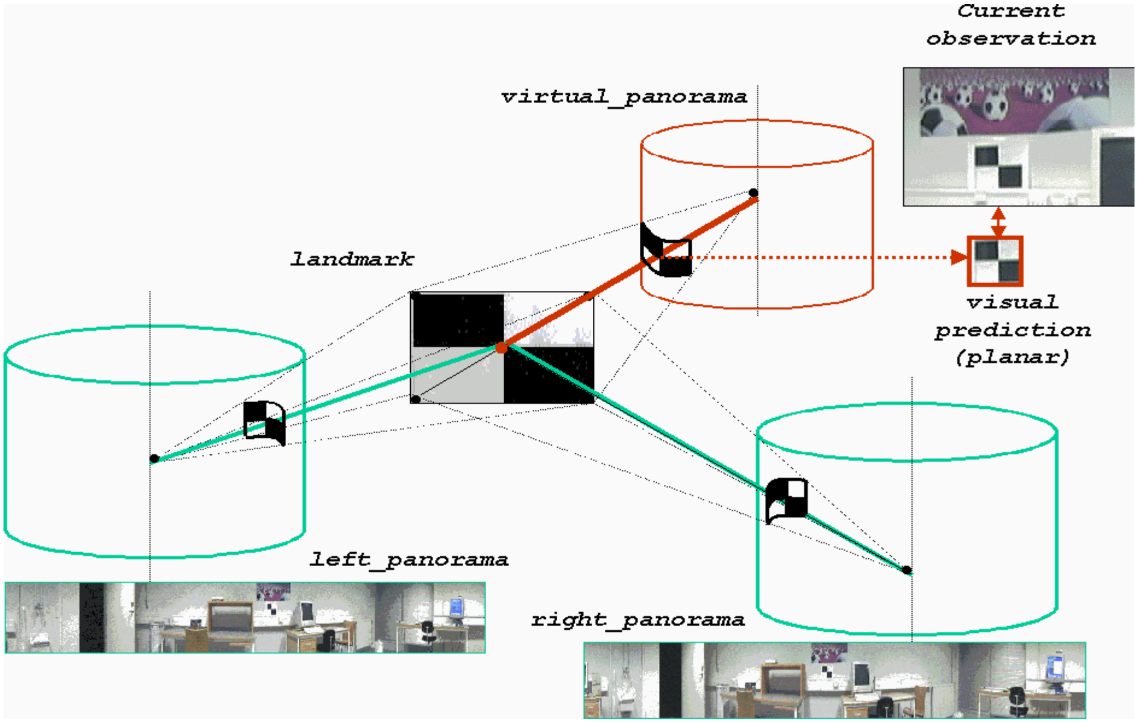


Figure 6.6: The "transfer of texture" from cylindrical reference-views to the visual prediction (see text).

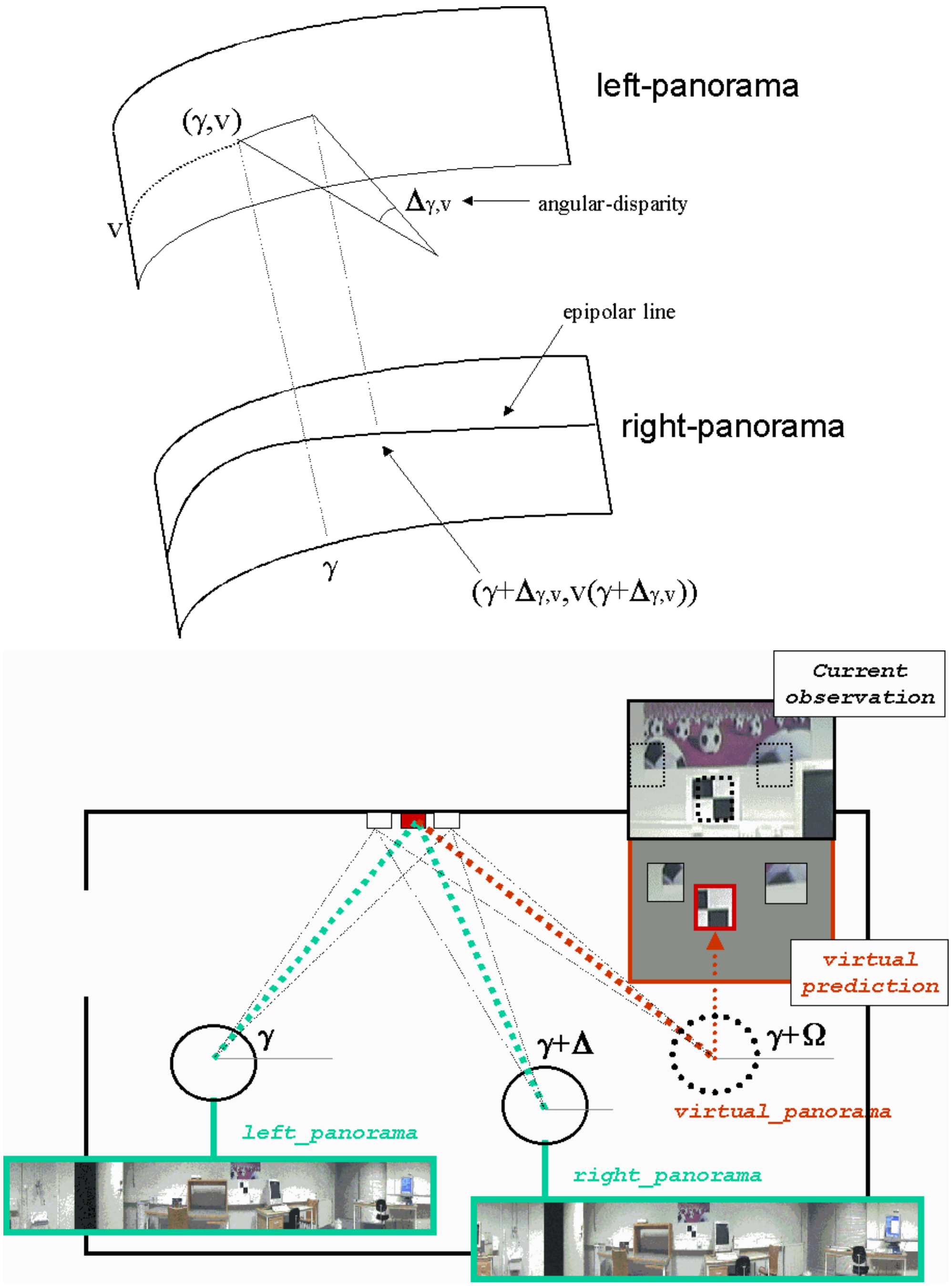


Figure 6.7: The top figure illustrates pixel correspondences in two different cylindrical panoramas and the related *angular disparity*. The bottom figure illustrates the cylindrical-to-cylindrical mapping based on angular disparity through an example (which consider a workspace floor-map). The bottom figure also includes images of reference panoramas, visual predictions and current observation.

The angular disparity can be estimated in different ways depending on the available knowledge. For example, by manually or automatically specifying a sparse set of corresponding points that are visible in both reference-views, by knowing or recovering camera internal parameters, and by exploiting epipolar geometry, a dense set of corresponding points can be recovered. Examples of procedures which exploit epipolar relations to recover dense correspondences from a sparse set of corresponding points, can be found in different literature works, (McMillan and Bishop [103], Faugeras [51], Blanc, Livatino and Mohr [15], [14]).

In our case, the angular disparity is inferred from previously estimated correspondences along cylindrical epipolar lines. These correspondences allowed for estimating the geometry of the landmark surface based on: 3D positions of landmark center, a minimum of two landmark corners, and the result of the planarity test, (this is discussed in section 5.2). Landmarks are in our case required to lie on one planar or "almost planar" surface, nevertheless, the same procedure could be applied to landmarks lying on more than one surface (in case the surfaces are known).

The proposed rendering system takes as input cylindrical reference-views of landmarks, along with the map of the angular disparities. This information is used to automatically generate image warps that map landmark reference-views to arbitrary cylindrical landmark-views. Note that the generated warps are capable of describing perspective effects, and occlusions (using a simple visibility algorithm that guarantee back-to-front ordering [103]).

The cylindrical-to-cylindrical mapping is illustrated in figure 6.7. Each angular disparity value, $\Delta_{\gamma,v}$, can be obtained as in equation 6.1. Note that (γ, v) are the pixel coordinates in the panorama, where γ is an angle while v is the pixel row.

$$\Delta_{(W_\gamma^{left}, W_v^{left})} = W_\gamma^{right} - W_\gamma^{left} \tag{6.1}$$

where $(W_\gamma^{left}, W_v^{left})$ represents a generic pixel in the left cylindrical reference-view identified by the angle γ and the ordinate v, and $(W_\gamma^{right}, W_v^{right})$ represents the correspondent ordinate v for a certain angle γ, in the right reference-view.

Knowing the angular disparity for each landmark pixel, this can be converted for each position on the left cylinder $(\gamma,\ v)$, into an image flow vector field, $(\gamma + \Delta_{\gamma,v},\ v(\gamma + \Delta_{\gamma,v}))$. The figure 6.7 top row illustrates this conversion.

The disparity values can then be transfered from the known cylindrical pairs $(C_x^{left}, C_y^{left}, C_z^{left})$ and $(C_x^{right}, C_y^{right}, C_z^{right})$ (which respectively represent the left and the right camera positions), to a new cylindrical projection in an arbitrary position, $(C_x^{virt}, C_y^{virt}, C_z^{virt})$, using the following equations, where τ is the rotation offset which aligns the angular orientation of the cylinders to a common frame,

$$\begin{aligned}
a &= (C_x^{right} - C_x^{virt}) \cos(\tau - W_\gamma^{left}) + (C_y^{right} - C_y^{virt}) \sin(\tau - W_\gamma^{left}) \\
b &= (C_y^{right} - C_y^{left}) \cos(\tau - W_\gamma^{left}) + (C_x^{right} - C_x^{left}) \sin(\tau - W_\gamma^{left}) \\
c &= (C_y^{virt} - C_y^{left}) \cos(\tau - W_\gamma^{left}) + (C_x^{virt} - C_x^{left}) \sin(\tau - W_\gamma^{left})
\end{aligned} \tag{6.2}$$

$$\cot \Omega_{(W_\gamma^{virt}, W_v^{virt})} = \frac{a + b\ \cot \Delta_{(W_\gamma^{left}, W_v^{left})}}{c} \tag{6.3}$$

The resulting $\Omega_{(W_\gamma^{virt},W_v^{virt})}$ is the angular disparity between the generic pixel in the left cylindrical reference-view, $W_{\gamma,v}^{left}$, and the corresponding pixel in the virtual cylindrical view. In this way, each resulting angular disparity value, $\Omega_{\gamma,v}$, can be converted, for each position on the left cylinder $(\gamma,\, v)$, into an image flow vector field $(\gamma + \Omega_{\gamma,v},\, v(\gamma + \Omega_{\gamma,v}))$ using the epipolar relation given by equation 6.4, (i.e. the same as equation 5.10 in chapter 5), but in this case it is applied to the virtual cylinder.

$$W_v^{virt}(W_\gamma^{virt}) = \frac{M_x \cos(\tau - W_\gamma^{virt}) + M_y \sin(\tau - W_\gamma^{virt})}{M_z} + C_v \tag{6.4}$$

where

$$\begin{bmatrix} M_x \\ M_y \\ M_z \end{bmatrix} = \left(\begin{bmatrix} C_x^{left} \\ C_y^{left} \\ C_z^{left} \end{bmatrix} - \begin{bmatrix} C_x^{virt} \\ C_y^{virt} \\ C_z^{virt} \end{bmatrix} \right) \times \begin{bmatrix} \cos(\tau - W_\gamma^{left}) \\ \sin(\tau - W_\gamma^{left}) \\ C_v - W_v^{left} \end{bmatrix} \tag{6.5}$$

and

$$W_\gamma^{virt} = W_\gamma^{left} + \Omega_{\gamma,v} \tag{6.6}$$

where τ is the rotation offset which aligns the angular orientation of the cylinders to a common frame, and C_v is ordinate v of the scan-line where the center of the projection would project onto the scene, (i.e. the ordinate of the line of zero elevation).

The above equation gives a concise expression for the curve, $W_v^{virt}(W_\gamma^{virt})$, (i.e. the cylindrical epipolar line), formed by the projection of a ray across the surface of a cylinder, (labeled "virt"), where the ray is specified by its positions on some other cylinder, (labeled "left").

Once the angular disparity, $\Delta_{\gamma,v}$, has been used for the transfer of the disparity values between the reference cylinder to a new viewing position, each estimated pixel in the virtual cylinder, $W_{i,j}^{virt}$, is projected on the virtual camera image-plane, so becoming $W_{s,t}^{virt-plan}$, in order to generate the landmark visual prediction. The visual prediction is converted to planar to be compared to landmark current observation. Figure 6.8 illustrates the mapping from cylindrical to planar image.

The proposed landmark pixel transfer procedure starts with a forward mapping (from reference to virtual) for what concern the landmark corners (which position has been previously established by stereo-matching, see section 5.2). This results in a region of the virtual-cylinder delimited by the four projected corners. Each pixel included in the delimited region is then projected forward or inverse, following the consideration made in sub-section 4.1.2. In particular, since our aim is a high texture fidelity, a forward mapping is adopted in the compression case and an inverse mapping in the enlargement case.

In summary, the proposed technique of cylindrical pixel transfer allows for establishing pixel correspondence between landmark reference-views and virtual prediction. In particular, two reference views are considered in our case . Experiments showed that the proposed technique allows for reliable matches between prediction and observation even in presence of significant positional errors, (denoted by clear displacements in image-plane between prediction and observation). See experimentation section for more details.

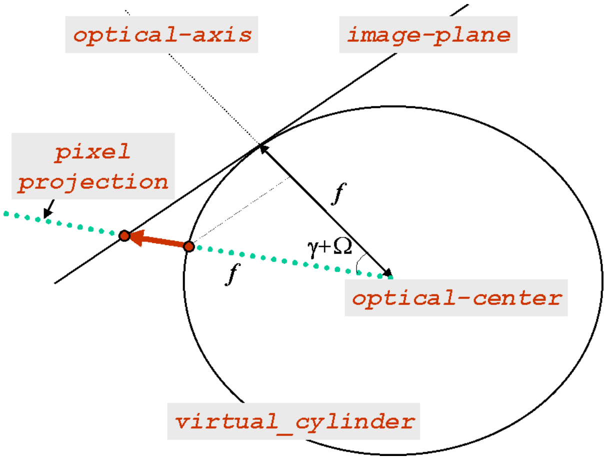


Figure 6.8: The figure illustrates the mapping from cylindrical to planar image.

6.3.2 View-Adapted Texture Mapping

Once a correspondence between landmark reference-views and visual prediction has been established, the texture values in the reference-views can be mapped to the virtual view. It is proposed to apply projective texture mapping in an *adaptive* way, in order to take advantage of multiple reference-views of landmarks, (from different positions), depending on the current robot pose.

The basic concept is that the appearance of an object in the camera image-plane strongly depends on relative position between camera and object, and on present light conditions. In particular, different views of an object may reveal different visible aspects, (e.g. disocculsions), and local illumination effects, (e.g. shadows, reflections, highlights, etc.). Figure 6.9 summarizes the main factors affecting appearance of objects when observed from different viewpoints, (the figure represents a part of the scheme in figure 5.1).

But, how should texture values of different reference-views contribute when transfered to the same pixel in the virtual view?

In some of the literature works related to realistic visualization the issue of blending multiple images have been addressed, and it has been demonstrated the advantage of considering more than one reference texture when generating texture on a virtual view.

Different techniques have been proposed for blending texture values relative to different views. Among them, simple weighting functions based on the angle of the camera to the object, to more sophisticated post-rendering calculations, (Mark et al [99]). In case a geometric model is available for the represented objects, (even if this is a coarse model), textures could efficiently be mapped by a view-dependent projective mapping as shown in Debevec, Yu and Borshukov [44]. In Debevec, Taylor and Malik [43] it is shown that such a mapping could also be exploited to refine the geometric model of an object by a technique named: *model-based stereo.*

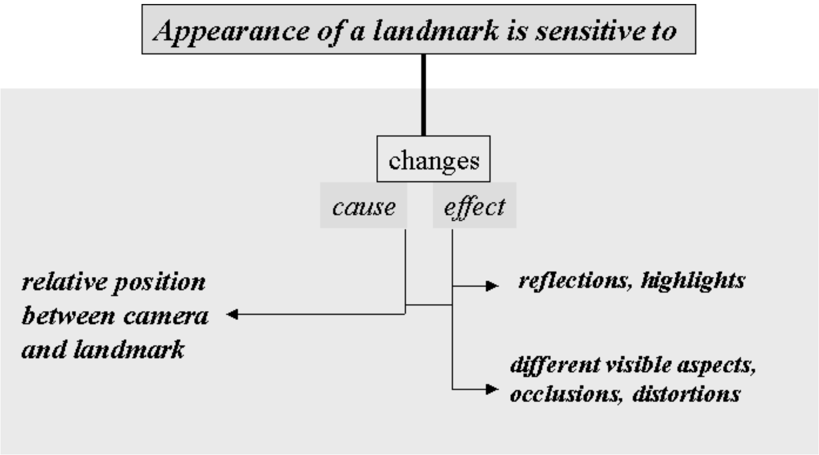

Appearance of a landmark is sensitive to
changes
cause
effect
relative position between camera and landmark
reflections, highlights
different visible aspects, occlusions, distortions

Figure 6.9: The figure summarizes main factors affecting appearance of objects when observed from different viewpoints.

In case of proposed system, it is proposed to merge two o more landmark reference views into a composite rendering, combining texture values of correspondent pixels in reference views. In particular, the system calculates a weighted average where involved textures provide a different contribution. Images from reference views which are closer to current viewpoint are expected to better approximate the current view than a reference-view further away. Closer images are in fact expected to best approximate landmark visible aspects and local illumination effects present in current camera observation.

Figure 6.11 shows an example situation which can also be referred to the landmark of figure 6.10 (a portion of the computer monitor). In case of a planar object with a planar neighbor region, it is not that relevant which reference view should provide a higher contribution when estimating the texture of the current view. In case of an "almost" planar object with a not planar neighbor region (as the case of the monitor), the closest reference view (ref_1 in figure 6.10) should provide the higher contribution. In fact, the closest view contains "reflections" and visible aspects not shown in the farer reference view (ref_2).

In particular, our system calculates a weighted average where involved textures provide a different contribution which depends on:

- *the magnitude of the angle* between the lines that connect the landmark center with the optical centers of the camera, (related to considered reference and current view). This angle is depicted in figure 6.10 as α_i , $i = 1, 2$. The weight for this angle is inversely proportional to the magnitude of the angle.
- *the distance* between the current viewing position and the reference positions. This distance is depicted in figure 6.10 as d_i , $i = 1, 2$. The weight for this distance is inversely proportional to the length.

The reference view which is closer to current viewpoint in "angle" and in "distance" will thus give a higher contribution in the final summation.

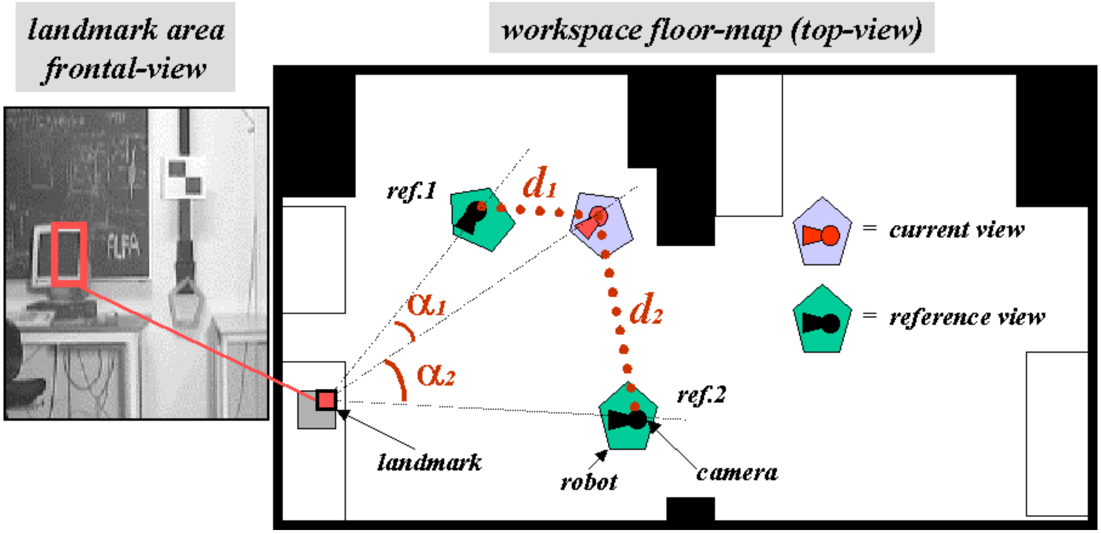


Figure 6.10: The figure left hand side shows the considered landmark (representing a part of the computer monitor). The figure right-hand side represents the angles between the lines that through the landmark center intersect the camera optical-center (angles α_1 and α_2), and the distances between current viewing position and the reference positions.

In case of two reference views, the resulting texture value for a pixel, VP, would then be calculated as in the following.

$$VP_{\alpha} = Ref_1 \ \frac{\alpha_2}{\alpha_1 + \alpha_2} + Ref_2 \ \frac{\alpha_1}{\alpha_1 + \alpha_2} \tag{6.7}$$

$$VP_{dist} = Ref_1 \ \frac{d_2}{d_1 + d_2} + Ref_2 \ \frac{d_1}{d_1 + d_2} \tag{6.8}$$

$$VP = \frac{VP_{\alpha} + VP_{dist}}{2} \tag{6.9}$$

The above equations can naturally be extended to the case of more than two reference views.

Merging reference-views based only on the above criteria can cause visible seams in the landmark visual prediction due to specularity, and unmodeled geometric detail may arise when neighboring textures comes from different reference-images and in case of occlusions or "disocclusions". Some of the techniques proposed in the literature for calculating texture transitions between different mapped views could then be applied to cope with the problem.

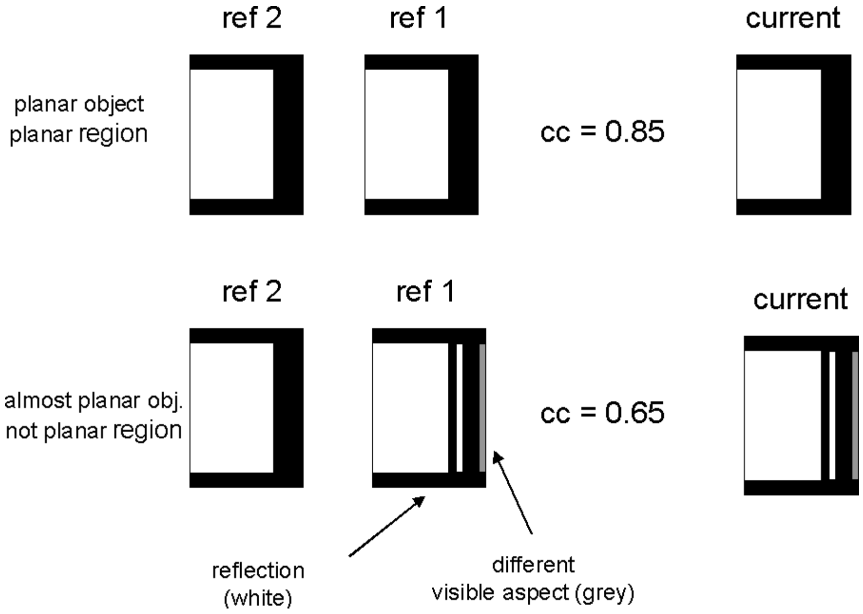


Figure 6.11: Figure shows an example situation, which can also be referred to the landmark of figure 6.10, where the reference view ref_1, which is the closest to current view, is expected to provide a better texture information than the reference view ref_2 which is farer. Possible reflections, highlights, shadows, and visible aspects, contained in the closest reference view, best approximate the current view texture. The figures top-row represent a portion of a planar object with a planar neighbor region, while the figures bottom-row represent a portion of an "almost" planar object, (e.g. the computer monitor), with a not planar neighbor region. "cc" represents an example of correlation coefficient which may result when matching the ref_1 and ref_2 views.

In the context of proposed method to robot navigation, main reasons for proposing view-adapted texture mapping can be summarized in:

1. a more reliable match between visual prediction and current observation;
2. an advantageous way of blending landmark reference-views when current viewpoint encompasses an area not entirely observed in one of the reference-views.

6.4 Experimentation

This section describes the experimentation involving automatic recognition of landmarks. The goal was to test performance of the proposed method, involving cylindrical pixel transfer and view-adapted texture mapping, related to the case when landmark visual and positional information have been automatically learned by the system. In particular, it was believed important to experiment with the following in a realistic setting:

- Accuracy and Reliability of Proposed Recognition Method

  a) transfer of textures in generated visual prediction;
  b) accuracy of matching algorithm;
  c) performance differences to previous method for visual prediction generation;
  d) match sensitivity to observation pose;
  e) match sensitivity to positional error;
  f) match sensitivity to texture structure;

- Extent and Shape of Landmark Observation Area.

  g) observation area sensitivity to landmark orientation;
  h) observation area sensitivity to texture structure and positional error.

The proposed method was implemented and tested on the mobile robotic system presented in previous chapters (section 4.1.5). The workspace for the experiments was the new laboratory (also described in section 4.1.5), whose floor-map is depicted in 6.12.

The experimentation was entirely performed in the new-laboratory workspace which contained a large area free of furniture, so suitable for testing landmark observation-areas expected larger than those experienced during the first experimentation phase in the old laboratory, (section 4.1). Unfortunately, the new laboratory workspace also contained very few textures which potentially challenged the acquisition of discriminant texture patches. However, few poster-pictures with various textures were introduced in the laboratory-room in order to provide the environment with more typical interior characteristics. The previously described "ideal" landmark, (i.e. textured as in figure 6.12), was also introduced in order to test discrepancies in performance between ideal and typical landmark texture-structures.

In order to test visual prediction reliability and matching performance, the system was asked first to self-learn landmarks and then automatically recognize the learned landmarks. The experimentation consequently required a number of learning runs followed by a number of recognition runs. During the recognition runs, the robot was several time set to execute a recognition process based on previously self-learned landmarks.

Every time the system is asked to recognize a landmark, it computed the most convenient observation angle from current robot position. This angle was defined as the angle occurring when the camera optical axis passes through landmark center. The angle is computed based on pose-estimates of camera, robot, and landmarks. The camera was then panned according to the computed angle before the image was captured. Figure 6.13 shows the *"most convenient"* observation angles for four different camera poses.

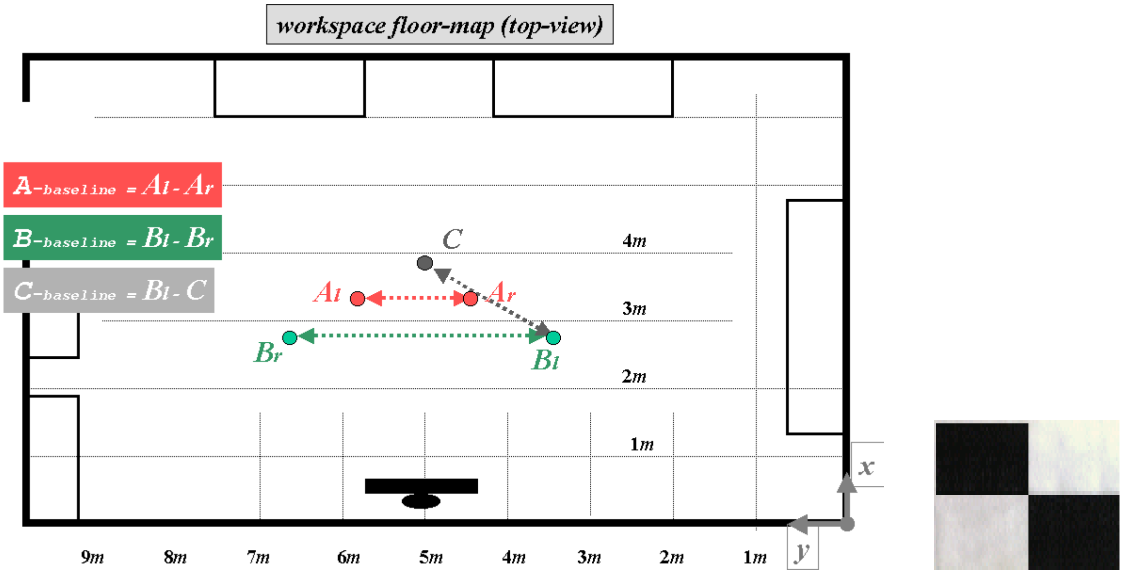


Figure 6.12: The left-hand figure shows the workspace for the experiments and the baselines of the learning trajectories. The right-hand figure shows the ideal landmark introduced in the environment in order to test discrepancies in performance between ideal and typical texture-patches

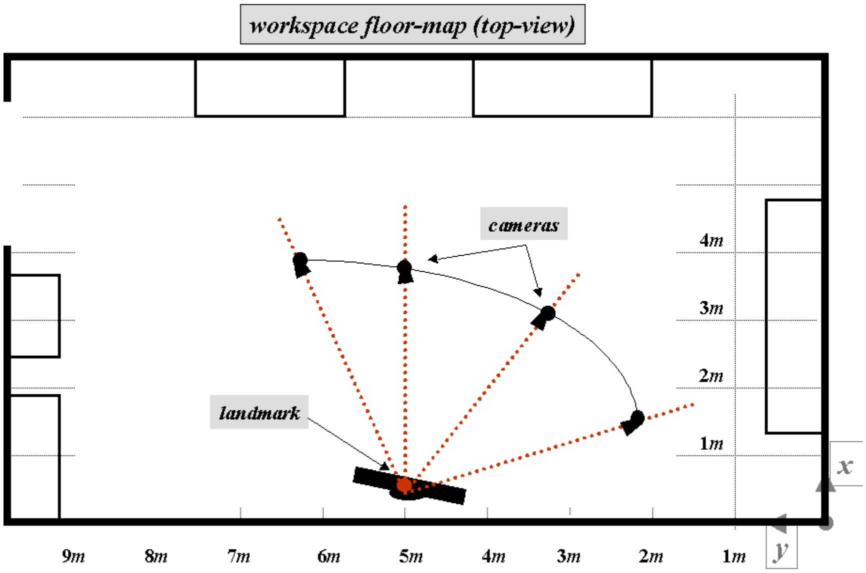


Figure 6.13: The figure shows a workspace floor-map representing the "most convenient" observation angle for four different camera poses and the same landmark. This angle is defined as the angle occurring when the camera optical axis passes through landmark center.

Camera pose at initial time and at each succeeding stop was manually and precisely measured by techniques based either on tape-meter measurements or on observation of three specialized landmarks (as explained in sub-section 5.2.4). On top of measured poses, an error was added with the intention to replicate naturally occurring positional inaccuracies. The error, (approximately 4 cm. in average), was randomly generated and based on realistic estimates. That is, similar to those obtained in previous work performed by the author, (Livatino and Madsen [93]), where the robot was let to move autonomously under the control of the self-localization algorithm described in chapter 4.

6.4.1 Accuracy and Reliability of Proposed Recognition Method

The purpose of this experimentation was to test the accuracy and reliability of the proposed recognition method based on the analysis of the correlation coefficient and discrepancies between estimated and observed landmark locations in the camera image-plane (in current observations).

The tests involved landmarks having different textures, poses, and positional errors. The tests mostly involved the recognition method proposed in this section, (related to automatically learned landmarks). In some of the recognition tests, the previously proposed recognition method, based on manual training and on one optimal reference-view, was applied for comparison.

Because of the many parameter involved, (mainly landmark texture, pose, positional error and recognition method), are strictly correlated to each other, it is impossible to experiment with them separately one at a time. Consequently, the results of typical runs will first be presented and afterwards the results will be analyzed by topic of investigation.

The experimentation was designed as described in the following.

- **Learning Phase**

  During different runs of the learning phase the robot was asked to follow different trajectories. These trajectories included two positions where the robot was stopped for the acquisition of a panoramic-view. In particular, the learning trajectory t^{learnA}, represented in figure 6.14 top-row, included the learning positions A^{right} and A^{left}, the trajectory t^{learnB}, represented in figure 6.14 bottom-row, included the positions B^{left} and B^{right}, and the trajectory t^{learnC} included the positions B^{left} and C.

  The traveled distance between two consecutive learning positions represented the baseline of the acquisition stereo-configuration. The baseline ranged from 0.97 meters of the $A^{baseline} =\mid A^{left} - A^{right} \mid$, to 1.34 meters of the $C^{baseline} =\mid B^{left} - C \mid$, to 2.93 meters of the $B^{baseline} = B^{left} - B^{right}$. The baselines of the learning trajectories are represented both in figures 6.12 and 6.14.

  From each of the learning positions the system acquired panoramic images and then processed sub-portions of them (i.e. sub-panoramas). Figure 6.16 shows a typical example of a sub-panoramic image including extracted texture-patches, surrounded by a white rectangle. When processing the same sub-panorama with different attention parameters an average of 13 landmarks were extracted at each run, and among them 3 or 4 landmarks were elected. The elected texture-patches for one of these runs are

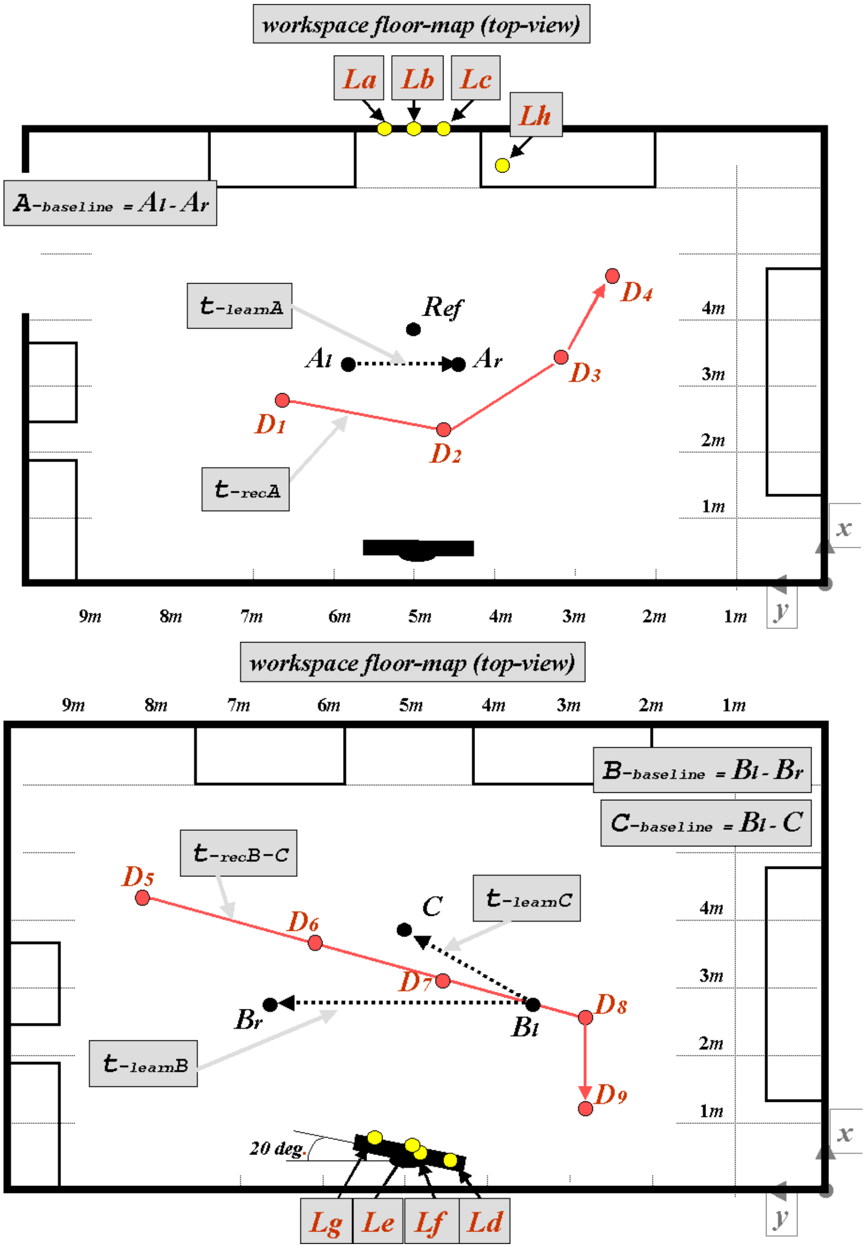


Figure 6.14: The figures show the new-laboratory workspace floor-map. In particular, the figures illustrate: landmark position (L_j, *j=a,..h*), learning trajectories (*t-learnA* and *t-learnB*), recognition trajectories (*t-recA* and *t-recB-C*), acquisition baselines (A_l-A_r, B_l-B_r, C-B_r), recognition position (D_i, *i=1,..,9*).

denoted L_a, L_b, L_c and L_h. The estimated landmark 3D positions, orientations, and measured errors of elected landmarks are shown in table 6.1.

Figure 6.18 shows a typical example of another sub-panoramic image including extracted texture-patches. When processing sub-panoramas with different attention parameters, an average of 15 landmarks were extracted in this case at each run, and among them 4 or 5 landmarks were elected. The elected texture-patches for one of these runs are denoted L_d, L_e, L_f and L_g. The estimated landmark 3D positions, orientations, and measured errors of elected landmarks are shown in table 6.2.

In tables 6.1 and 6.2, the elected landmarks are classified on their texture structure, denoted as ideal, good, sufficient and poor, and on their positional error, denoted as high, medium, low, minimum. The positional error is also classified by a number corresponding to the summation of errors in each of the coordinate axis.

- **Recognition Phase**

  In the recognitions runs, the robot was asked to follow different trajectories. These trajectories included several positions where the robot was stopped and the proposed automatic recognition method was run with the purpose of identifying previously self-learned landmarks. In particular, the recognition trajectory t^{recA}, represented in figure 6.14 top-row, included the recognition positions denoted D_1 through D_4, and the trajectory t^{recB-C}, represented in figure 6.14 bottom-row, included the positions denoted D_5 through D_9. The recognition positions were chosen in order to represent normal and critical conditions, (a critical condition is for example the edge of an observation-area).

  The generated visual predictions were matched to observations using the normalized cross-correlation algorithm (see equation 4.2). The search-window size was set based on the criteria discussed in sub-section 4.1.3. The landmark recognition process required around 2 seconds to run with our hardware, mostly employed by the normalized cross-correlation process.

  The robot was stopped at each recognition position in order to measure robot and camera pose. Nevertheless, the robot does not in principle need to stop when observing landmarks, (as long as its speed does not cause motion blurred images and the on-board camera is not panning). This has been demonstrated in previous work performed by the author, (Livatino and Madsen [93]).

  Diagrams in figures 6.17 and 6.19 shows results of the recognition tests related to recognition positions D_i, $i = 1,..,9$, (shown in figure 6.14), and landmark texture-patches L_j, $j = a,..,h$, (shown in figures 6.16 and 6.18).

  For each of the recognition positions D_i, the resulting correlation coefficient is reported for the visual predictions estimated by the proposed recognition method (related to automatic learning), including the case when a different number of reference-views is used. In case of figure 6.17 the resulting correlation coefficient is also reported for visual predictions estimated by the method previously (section 4.1), related to manual training. In the latter, the manually learned most representative reference-view was taken from position Ref in figure 6.14 top-row.

  In case of figure 6.19 the resulting correlation coefficient is reported for a different number of reference-views. Furthermore, in case of landmarks denoted L_e and L_g, the result is reported for reference-views related to different baselines (and thus different positional errors), denoted $B - bas$ and $C - bas$.

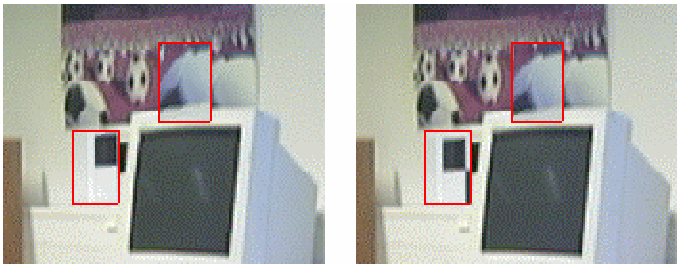


Figure 6.15: The figure shows examples of generated visual predictions superimposed to current observations. The left-hand side shows two visual predictions reconstructed based on one representative reference-view only, (previous method). The right-hand side shows visual predictions reconstructed by two reference-views, (proposed method). The reconstructed views in the right-hand side show a slightly larger wall portion.

The results of performed experiments are in the following analyzed according to the topic of investigation.

a) **Transfer of Textures in Generated Visual Predictions**

The first analysis concerns proposed cylindrical pixel transfer and view-adapted texture mapping.

Figures 6.15 and 6.25 shows examples of generated visual predictions superimposed to current observation for some of the recognition positions. Unfortunately, due to reduced dimensions of texture-patches and landmark simple geometries, it was sometime difficult to appreciate differences in accuracy among visual predictions reconstructed from different reference-views, by just looking at them. And this was also the case when predictions were reconstructed based on one representative reference-view manually learned, (previous method).

Nevertheless, from a first analysis based on "visible aspects", it is possible to state the following facts.

- The accuracy of textures in visual predictions generated by the proposed method was approximately the same as the previous method. For example in figure 6.15 the visual predictions at the right-hand, generated from two reference views automatically learned, approximately show the same accuracy as the visual predictions at the left-hand, generated from one reference-view manually learned. This demonstrates that pixel transfer and texture mapping were correctly and accurately computed.
- The visual predictions generated from two reference-views usually show a larger wall portion texture details (due to the different observation angle). For example in figure 6.15 the visual predictions represented in the right-hand shows more of the "ideal" landmark texture than the predictions in the left-hand.

The next sections demonstrate the improvement in visual prediction fidelity when using two reference-views, based on the analysis of the correlation coefficient.

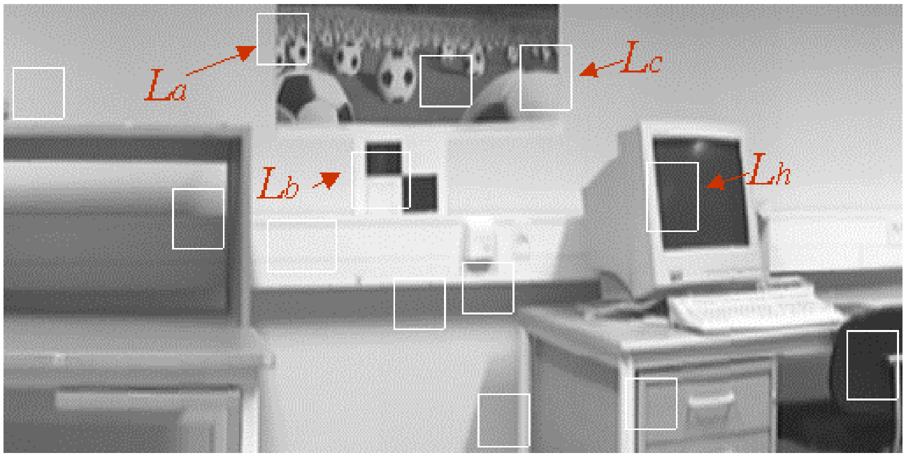


Figure 6.16: The figure shows a typical example of a sub-panoramic image including extracted texture-patches, surrounded by a white rectangle. The elected texture-patches are denoted L_a, L_b, L_c and L_h.

Landmark (texture-struct) (position error,sum)	cord.	Computed Value (mm)	Ground Truth (mm)	Measured Error (mm)
L_a	x	5755	5804	-49
(sufficient)	y	5462	5492	-30
(medium,101)	z	1426	1404	22
	γ	93.9 (deg.)	90.0 (deg.)	3.9 (deg.)
L_b	x	5843	5812	31
(ideal)	y	4943	4932	11
(low,59)	z	986	1003	-17
	γ	87.0 (deg.)	90.0 (deg.)	-3.0 (deg.)
L_c	x	5796	5776	20
(good)	y	4539	4536	3
(minimum,31)	z	1253	1245	8
	γ	90.9 (deg.)	90.0 (deg.)	0.9 (deg.)
L_h	x	5227	5334	-107
(sufficient)	y	4190	4238	-48
(high,208)	z	1006	953	53
	γ	102.9 (deg.)	96.0 (deg.)	6.9 (deg.)

Table 6.1: The table shows estimated landmark 3D positions, orientations, and measured errors of elected landmarks. The elected landmarks are classified on their texture structure, denoted as *ideal*, *good*, *sufficient* and *poor*, and on their positional error, denoted as *high*, *medium*, *low*, *minimum*. The positional error is also classified by a number corresponding to the summation of errors in each of the coordinate axis.

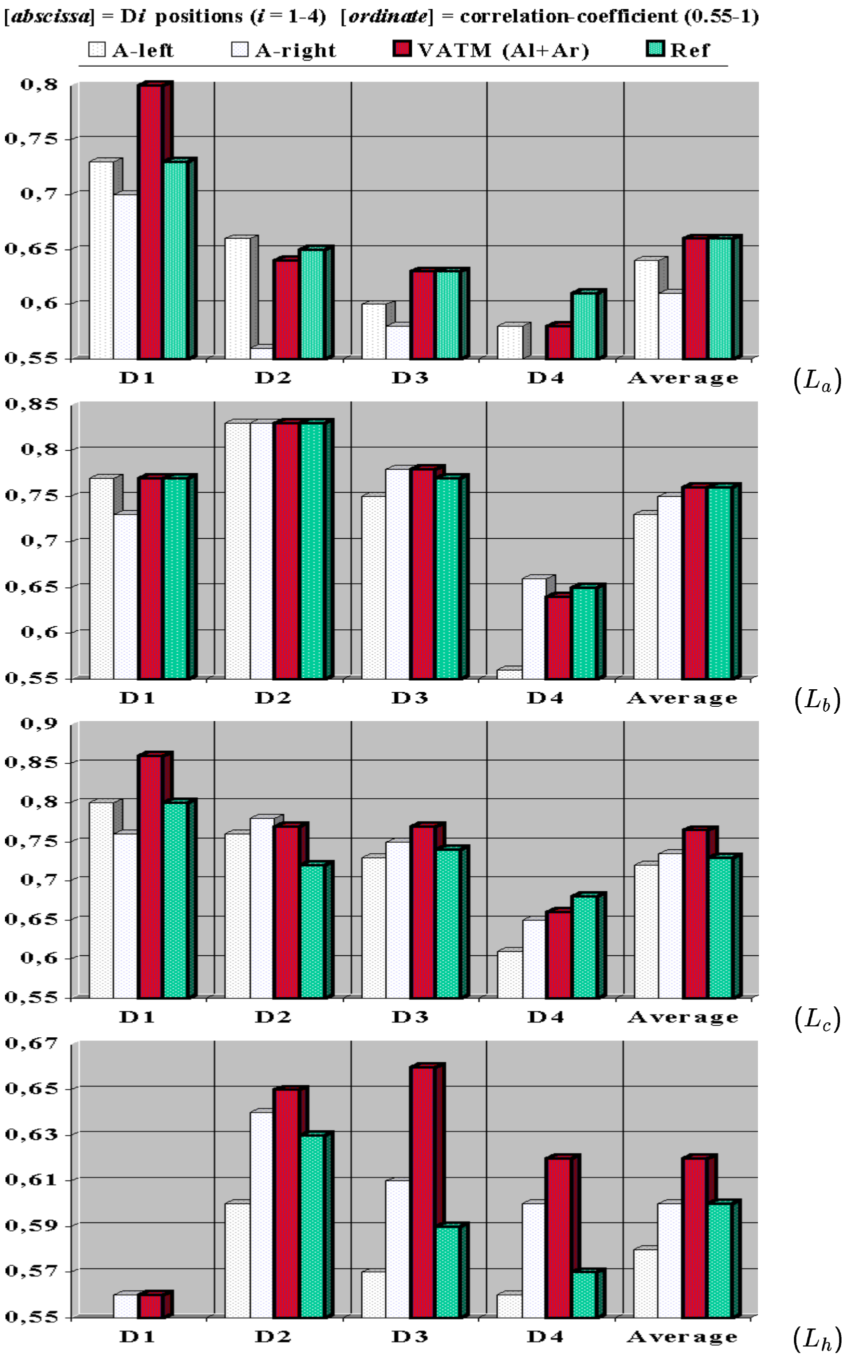


Figure 6.17: The diagrams in figures show results of the recognition tests related to recognition positions D_i, *i=1,..,4*, (shown in figure 6.14), and landmark texture-patches L_j, *j=a,b,c,h*, (shown in figure 6.16). The first "column" on the left (in grey) reports the response of the match only based on the left reference-view *A-left*. Analogously for the *A-right*. The red "column" represents the response of the proposed view-adapted texture mapping (VATM), while the green "column" represents the response of the match only based on the "ref" reference view.

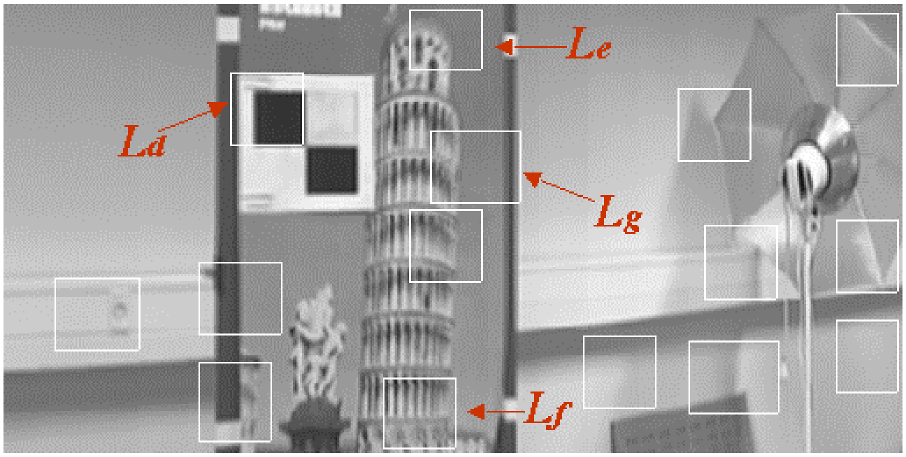


Figure 6.18: The figure shows a typical example of a sub-panoramic image including extracted texture-patches, surrounded by a white rectangle. The elected texture-patches are denoted L_d, L_e, L_f and L_g.

b) Accuracy of Matching Algorithm

This analysis concerns the accuracy of estimated landmark locations in current camera image-plane for successful matches.

The images in figure 6.15 and 6.25 show an error in predicted landmark location. The landmark visual prediction is in fact displaced to the observation. This displacement mainly arises from an error in robot pose and/or in estimated landmark pose, leading to the fact that landmark location is not where expected. The normalized cross-correlation process allowed for the recover of this error, (as long as visual prediction falls inside the search-window), so that landmarks could precisely be identified in the observation image-plane. In particular, the results showed that even in presence of an input positional errors of the order of 4-5 centimeters both in landmark and robot pose, leading to clear displacements between predicted and observed locations, reconstructed textures led to reliable matches, (i.e. an average error of 2 pixels between estimated and measured landmark image-location, such as estimated in sub-section 4.1.5).

It is then possible to summarize that accuracy of estimated landmark location in the image-plane is not worse than with the previous method. Next paragraphs demonstrate this aspect based on the comparison of correlation coefficients for the different recognition methods.

Landmark	cord.	Computed Value	Ground Truth	Measured Error
(texture-struct.)		(mm)	(mm)	(mm)
(position error,sum)				
L_d	x	643	622	21
(ideal)	y	4638	4648	-3
(minimum,31)	z	1202	1195	7
	γ	78.8 (deg.)	80.0 (deg.)	-1.2 (deg.)
L_e B-baseline	x	820	802	18
(good)	y	4952	4980	-28
(low,61)	z	1318	1333	-15
	γ	82.4 (deg.)	80.0 (deg.)	2.4 (deg.)
L_e C-baseline	x	853	802	41
(good)	y	4990	4980	10
(low,69)	z	1316	1333	-17
	γ	76.1 (deg.)	80.0 (deg.)	-3.9 (deg.)
L_f	x	753	772	-19
(poor)	y	4952	4911	41
(low,71)	z	724	713	11
	γ	83.4 (deg.)	80.0 (deg.)	3.4 (deg.)
L_g B-baseline	x	784	824	30
(good)	y	5014	5039	5
(low,57)	z	1129	1107	22
	γ	77.1 (deg.)	80.0 (deg.)	-2.9 (deg.)
L_g C-baseline	x	885	824	61
(good)	y	5018	5039	-21
(medium,109)	z	1080	1107	-27
	γ	84.7 (deg.)	80.0 (deg.)	4.7 (deg.)

Table 6.2: The table shows estimated landmark 3D positions, orientations, and measured errors of elected landmarks. The elected landmarks are classified on their texture structure, denoted as *ideal*, *good*, *sufficient* and *poor*, and on their positional error, denoted as *high*, *medium*, *low*, *minimum*. The positional error is also classified by a number corresponding to the summation of errors in each of the coordinate axis.

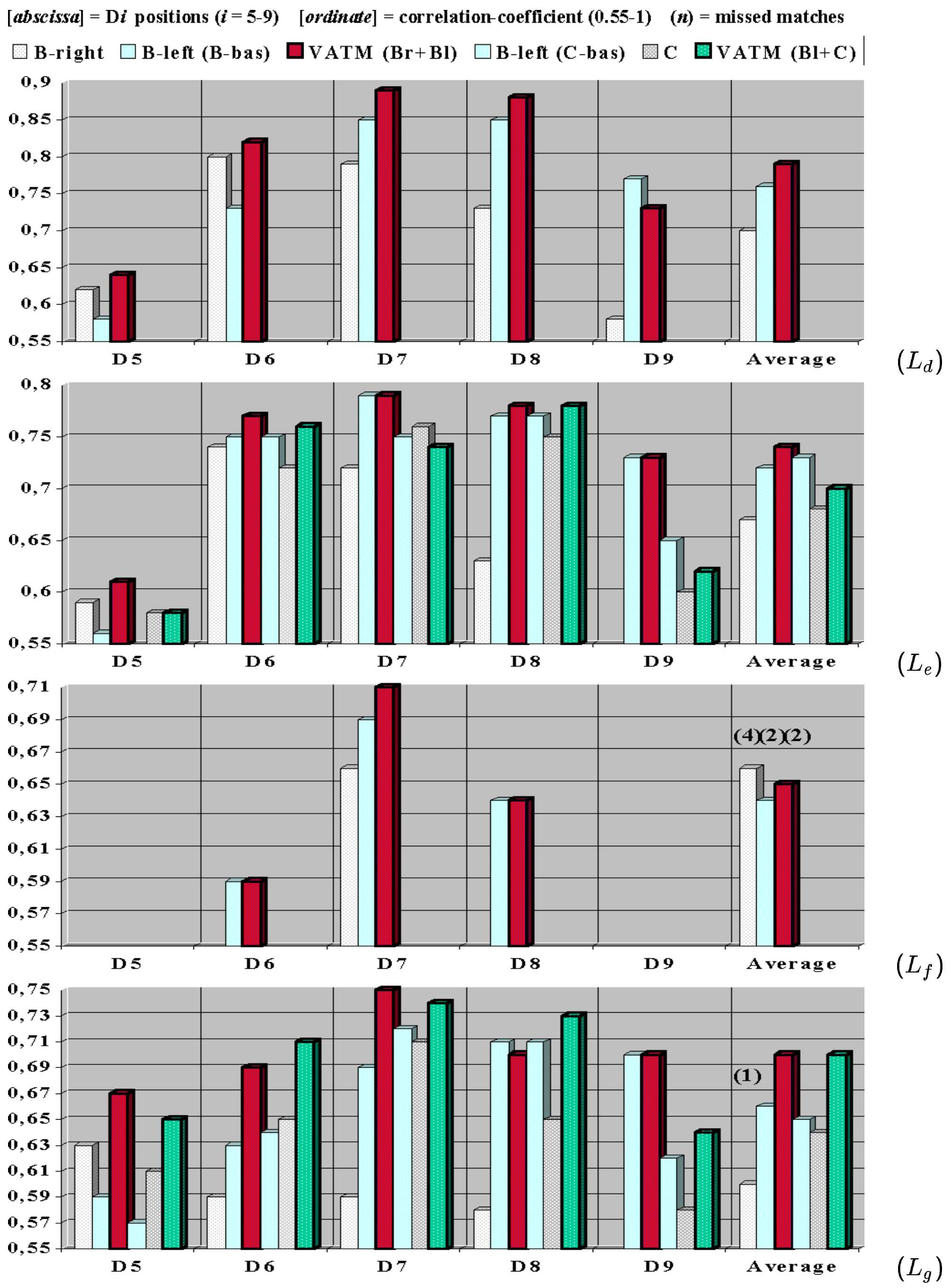


Figure 6.19: The diagrams in figures show results of the recognition tests related to recognition positions D_i, *i=5,..,9*, (shown in figure 6.14), and landmark texture-patches L_j, *j=d,e,f,g*, (shown in figure 6.18). The first "column" on the left (in grey) reports the response of the match only based on the left reference-view *B-right*. Analogously for *B-left (B- and C- baseline)* in light-blue, and *C* in dark-grey. The red "column" represents the response of the proposed view-adapted texture mapping (VATM) based on *B-right* and *B-left (B-baseline)*, while the green "column" represents the response of the VATM based on *B-left* and *C (C-baseline)*.

c) **Performance Differences to Previous Method for Visual Prediction Generation**

This analysis concerned with: (1) performance differences between the recognition method proposed in this sub-section for a different number of reference-views. In particular the considered reference-views were: the left-view only, the right-view only, both the right and the left views. (2) A comparison between the proposed recognition method and the one based on both manual training and one optimal reference-view previously proposed.

From diagrams in figure 6.17 and 6.19 the following facts can be stated.

- In most of the cases the normalized cross-correlation performed better, (i.e. higher correlation coefficient), when landmark visual predictions were generated based on two reference-views compared to the case when only one reference was used. For example the case of L_h-D_3 and L_c-D_1 in figure 6.17. The columns representing the average result well summarize this aspect. This showed the advantage of using the view-adapted texture-mapping.
- In few cases the prediction based on two references performed worse than the one based on one reference. For example, the case of L_a-D_2, L_b-D_4 in figure 6.17, and L_d-D_9 in figure 6.19. The main reasons seemed to be that either one of the reference views had much poorer texture than the other one, or the low positional accuracy affected the resulting "averaged" visual-prediction. Nevertheless, in both cases the proposed texture-mapping technique, (which inversely weights reference-views based on angle and distance to current viewing position), it is believed to "limit the damage".
- The advantage of using more reference-views and the view-adapted texture mapping is emphasized in cases of "almost planar" landmarks, (L_c, L_h, L_g). This confirmed the author's claims that the proposed method would show a dramatic improvement when applied to landmarks that are not planar.

The experiments show the view-adapted texture mapping performing better in approx. 70 % of the cases.

d) **Match Sensitivity to Observation Pose**

This analysis concerned with how the accuracy and reliability of the match may change according to different observation poses, which in turn means different distances and angles to landmarks. In particular, the distance between landmark-center and current camera optical-center, and the angle between landmark-plane and current camera image-plane.

From diagrams in figure 6.17 and 6.19, it is possible to summarize the following.

- Typically, a higher correlation coefficient was achieved when reference-views were closer to current observation. For example, in case of L_c-D_2, the prediction generated from Ref got the lower value because it was the most distant reference-view, in case of L_b-D_4, the predictions generated from A_r and Ref performed better than the one from A_l.
- Typically, a higher correlation coefficient was achieved when a reference-view had an angle to landmark closer to current angle to landmark. For example, in case of L_g-D_9, the prediction generated from B_l led to a successful match while the one generated from B_r did not provide enough data for a match.

- The recognition performance varied according to different distances and angles to landmark. In particular, above a certain distance or angle it was not possible to find any match. The magnitude of the critical angle and distance can roughly be quantified according to what stated in case of the previous method (see section 4.1.5). That is, the landmark texture should not be compressed to more than about a half of its original size for any of its dimensions, width and height, (however, for the most discriminant patches, landmark textures could be compressed to more), and landmark texture should not be enlarged to more than double of its original size. The critical angle in performed experiments ranged from approximately 30 to 60 degrees.
- Distance and angle to a landmark were not the only factor affecting matching performance. Figures 6.20 middle-row show the variation of correlation related to landmarks represented in figures 6.20 bottom-row (visual appearance) and figures 6.20 top-row (workspace position). Left-hand figures refer to landmarks L_a, L_b, L_c, L_h and positions D_1, D_2, D_3, D_4, while right-hand figures refer to L_d, L_e, L_f, L_g and positions D_5, D_6, D_7, D_8, D_9.

The experimental data however demonstrated, (based on the analysis of the correlation coefficient), a general advantage in weighting higher texture values related to reference views which are closer to current view, i.e. the advantage of weighting angle and distance to a landmark inversely proportional to their magnitude. This confirmed the choice pursued in proposed view-adapted texture mapping technique.

e) **Match Sensitivity to Positional Error**

This analysis concerns with how the accuracy and reliability of a match may change according to different errors in landmark position.

In order to easily refer to different errors affecting landmarks, the measured error was roughly categorized into four classes based on the summation of the absolute error in each of the coordinate axis. In particular, *minimum* refers to a summation result between 0 and 50 millimeters, *low* a result between 51 and 100, *medium* a result between 101 and 150, and *high* a result over 150. From diagrams in figure 6.17 and 6.19, it is possible to summarize the following.

- Typically, the higher the positional error the lower the correlation coefficient obtained from the match. This can be seen when comparing correlation response in figure 6.20 middle-row left-hand related to L_c, error class (minimum,31), and L_a, (error class medium,101), as well as when comparing L_b (low,59) and L_h (high,208). This can also be noted in case of a same landmark when the position has been estimated with a different error depending on the used stereo-baselines. The figure 6.21 top-row left-hand shows an example of this case, (landmark L_e). The estimate based on the $B^{baseline}$ is more accurate.
- The positional error is not the only factor affecting matching performance. The distance and angle to a landmark may in fact play a role as well. For example, in case of landmark L_g in figure 6.21 top-row right-hand, the match gives mostly similar correlation response despite the landmark pose estimates are characterized by a different error, (the estimate based on the $B^{baseline}$ is more accurate).

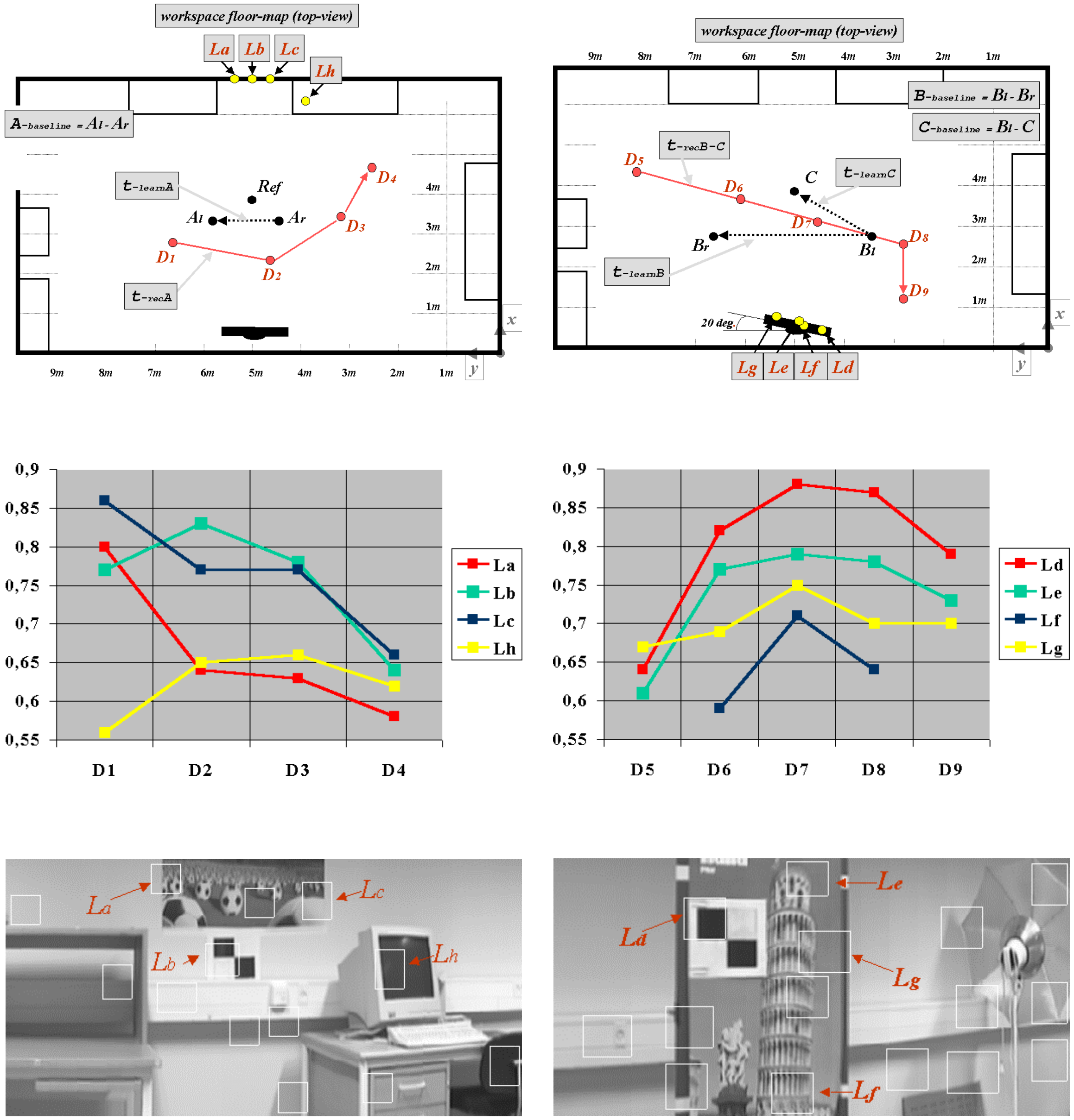


Figure 6.20: The figures middle-row show the variation of correlation related to landmarks represented in figures bottom-row (visual appearance) and figures top-row (workspace position). Left-hand figures refer to landmarks L_a, L_b, L_c, L_h and positions D_1, D_2, D_3, D_4, while right-hand figures refer to L_d, L_e, L_f, L_g and positions D_5, D_6, D_7, D_8, D_9.

In summary, the experiments demonstrated a better performance in landmarks which position was estimated with higher accuracy.

f) **Match Sensitivity to Texture Structure**

This analysis is concerned with how the accuracy and reliability of a match may change according to different texture structures.

In order to easily refer to different texture structures, landmark patches were roughly categorized into four classes based on the considerations made in chapter 4. In particular, *ideal* refers to high-contrasted textures with discriminant extended uniform regions, (for example a sub-portion of the "ideal" landmark), *good* refers to a more natural texture, less contrasted but still containing discriminant extended uniform regions, *sufficient* refers to a more ambiguous texture, and *poor* refers to a low-contrasted texture with limited uniform regions. From diagrams in figure 6.17 and 6.19, it is possible to summarize the following.

- Typically, the closer the texture structure is to ideal the higher the correlation coefficient obtained from the match. This behavior can easily be appreciated when comparing correlation responses related to landmarks with different texture-structures in figure 6.20 middle-row right-hand. The figure bottom-right represents correspondent landmark textures. This can also be noted in case of landmarks with approximately same positional error but different textures. The figure 6.21 bottom-row right-hand shows an example of this case, (landmarks L_e and L_f).
- Texture structure is not the only factor affecting matching performance. For example, in figure 6.20 mid-left landmarks L_b and L_c have mostly similar correlation response despite they are characterized by a different texture.

In summary, the experiments confirmed the advantage of high-contrasted patch containing discriminant extended uniform patterns both for specialized and more natural landmarks.

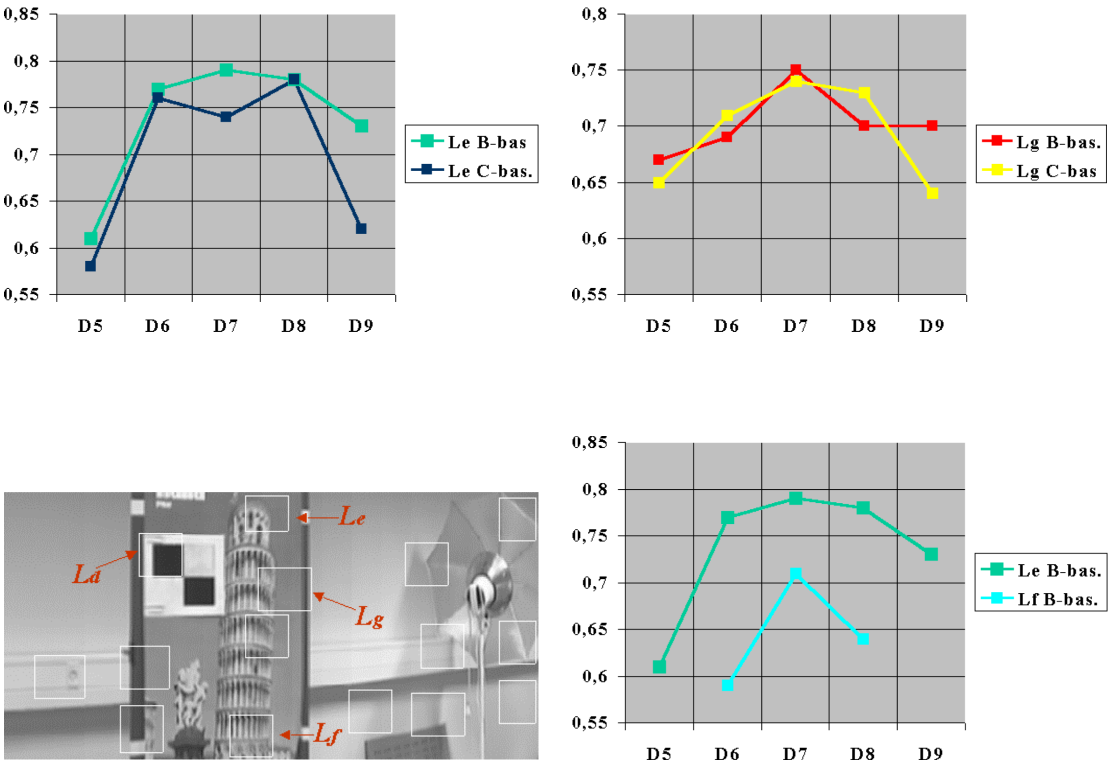


Figure 6.21: The figures show matching response (correlation coefficient) related to the landmarks represented in figure bottom-row left-hand. Figure top-row left-hand represents the case of the same landmark, L_e, which position has been estimated with a different error depending on the used stereo-baseline. The figure top-row right-hand shows the case when the match gives mostly similar correlation response despite the landmark poses have a different error, (landmark L_g). The figure bottom-row right-hand shows the case of landmarks with approximately the same positional error but different textures (landmarks L_e and L_f).

6.4.2 Extent and Shape of Landmark Observation Areas

The purpose of this experimentation was to test the extent and shape of observation areas, i.e. the workspace regions where a landmark can reliably be recognized, based on the response of the recognition process.

The tests involved landmarks having different texture-structures, poses, positional errors, and landmark angles. The tests only involved the recognition method related to automatically learned landmarks.

The experimentation was designed as described in the following.

- **Learning Phase**

  The learning trajectory considered for these experiments was the one denoted t^{learnB}, represented in figure 6.14 bottom-row, which included the positions denoted B^{left} and B^{right}, The baseline, denoted $B^{baseline}$, measured 2.93 meters.

  The learned landmarks were the ones denoted L_e and L_f in figure 6.18 plus a landmark representing the central area of the ideal landmark (i.e. landmark L_k in figure 6.24). The landmark estimated poses and measured errors were as described in table 6.2. The positional error for the landmark L_k was of the same order as for the one denoted L_e.

  The learned landmarks represented regions of a poster picture hanging on a rotating panel. A panel containing a poster on a steerable support was used in order to test observation areas for different landmark angles. Figure 6.23 left-hand shows the steerable support. The learning process was repeated every time the panel was set to a different angle.

- **Recognition Phase**

  In order to estimate the extent and shape of observation areas, the workspace was partitioned into a point grid. Figure 6.22 shows the workspace floor map containing the point grid. The figure shows a workspace floor-map.

  Each point in the grid represented workspace positions where the observed view was compared to the landmark visual-prediction. In particular, the robot was taken to each of the grid-points, its pose was measured, the camera was panned to achieve "most convenient" observation conditions, and eventually proposed recognition method was run using both the left and right reference-views previously learned.

  The extent and shape of the observation areas were determined based on the correlation response of the matches between visual-predictions and observations. A dark dot represents a successful match, i.e. correlation coefficient above threshold, while a white dot represents a missed match. The correlation threshold was set to 0.55. An example of how the observation area was determined in case of the second grid-line for landmark L_e is shown in figure 6.22.

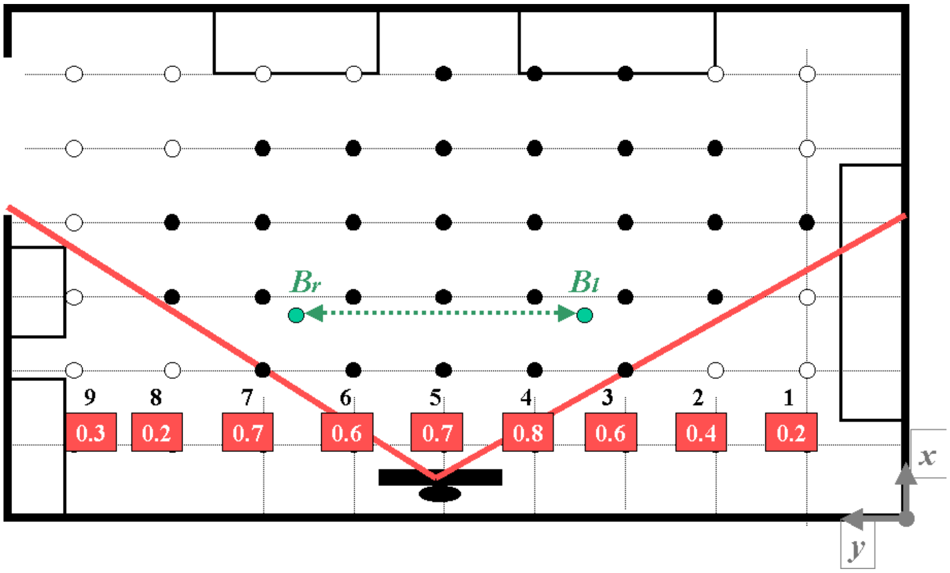


Figure 6.22: The figure shows how the observation area is determined in case of the second grid line from bottom. The values inside the red squares denote the correlation coefficients resulting from the match.

The results of performed experiments are in the following analyzed according to the topics of investigation.

g) **Observation Area Sensitivity to Landmark Orientation**

This analysis concerned with how the extend and shape of an observation area may change according to different landmark orientations. In particular, the angle between baseline and landmark plane. The sensitivity to landmark orientation was only tested for rotations around the vertical axis.

The rotating panel was turned to form an angles of a certain magnitude, then, observation areas were estimated. This test involved the landmark denoted L_e representing the tower top-right corner area as shown in figure 6.23 left-hand.

From the response of the tests it is possible to summarize the following.

- The recognition performance varied according to landmark angle. Figure 6.23 shows the case of how the extent and shape of the observation areas changed according to a different landmark angle. The angle between landmark plane and baseline were ranging from 0 to 20 degrees.
- A typical critical value for the landmark angle above which a landmark can not be recognized was 30 degrees.

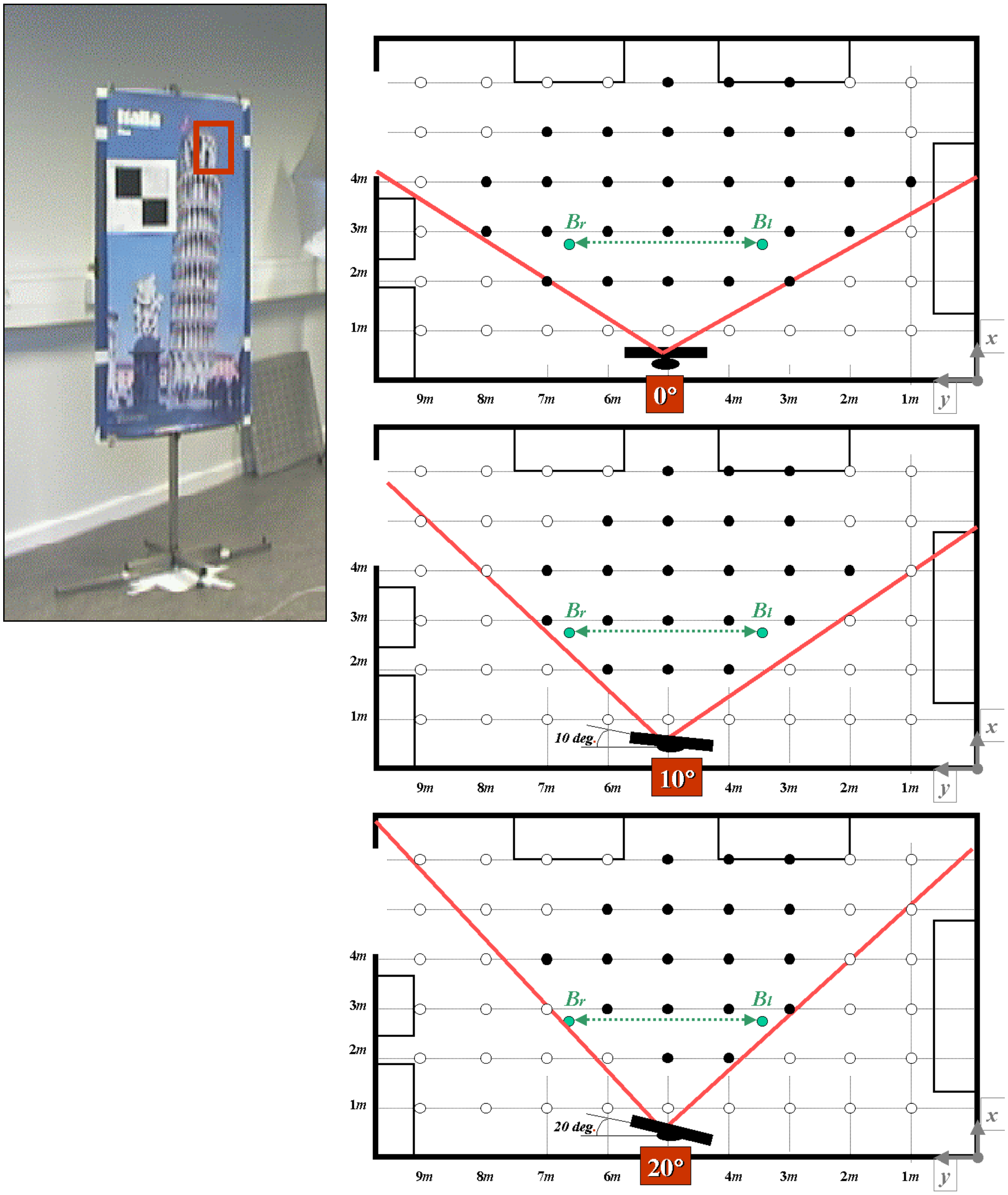


Figure 6.23: The figure left-hand side shows a panel on a steerable support introduced in order to test observation areas for different angles to landmark. The figures right-hand side show workspace floor-maps which describe observation areas related to the same landmark but with a different orientation.

h) **Observation Area sensitivity to Texture Structure and Positional Error**

This analysis concerned with how the extent and shape of the observation area may change according to different texture structures and positional errors.

This test involved the three landmarks described above, (see figure 6.24), represented by a different type of texture structure. From the response of the tests it is possible to summarize the following.

- The observation area varied according to landmark texture structure. In particular, textures closer to the ideal provide larger observation areas. Figure 6.24 shows computed observation areas for the three selected landmarks.
- The observation area varied according to landmark positional error. In particular, from the experience gained in performed experiments, it is possible to state that a larger error in landmark pose provides a smaller observation area. Nevertheless, in cases of landmarks denoted L_e and L_k a little variation in error poses (on the order of 1 cm), did not bring significant changes in recognition performance.
- An interesting issue was to observe that observation areas were not symmetric (as one could expect). In particular, the grid locations on the right-hand side of the figures representing the workspace, registered a higher number of matches. This seemed to be the consequence of a smaller positional error in camera B_l estimated pose.

In general, it is possible to summarize that the missed matches (white grid dots) were represented in most of the cases by either a current viewing position "too distant" from reference-view positions or by an angle between landmark surface and baseline "too large". In both cases the visual prediction had a small size and it could not be reliably matched to the observation.

A solution to this case was found by increasing the size of reference landmark-views. In fact, this led to increased size in visual predictions. The size of landmark reference views can only be increased under the assumption of planarity for a wide area around a selected landmark. This assumption can be made for areas containing a dense set of landmarks having same orientation. Despite an increased landmark size could not cope with the problem of an occlusion, this solution could be successfully used in many circumstances. Figures 6.25 right-hand show images of visual prediction where landmark size was increased. The figures left-hand represent the original size.

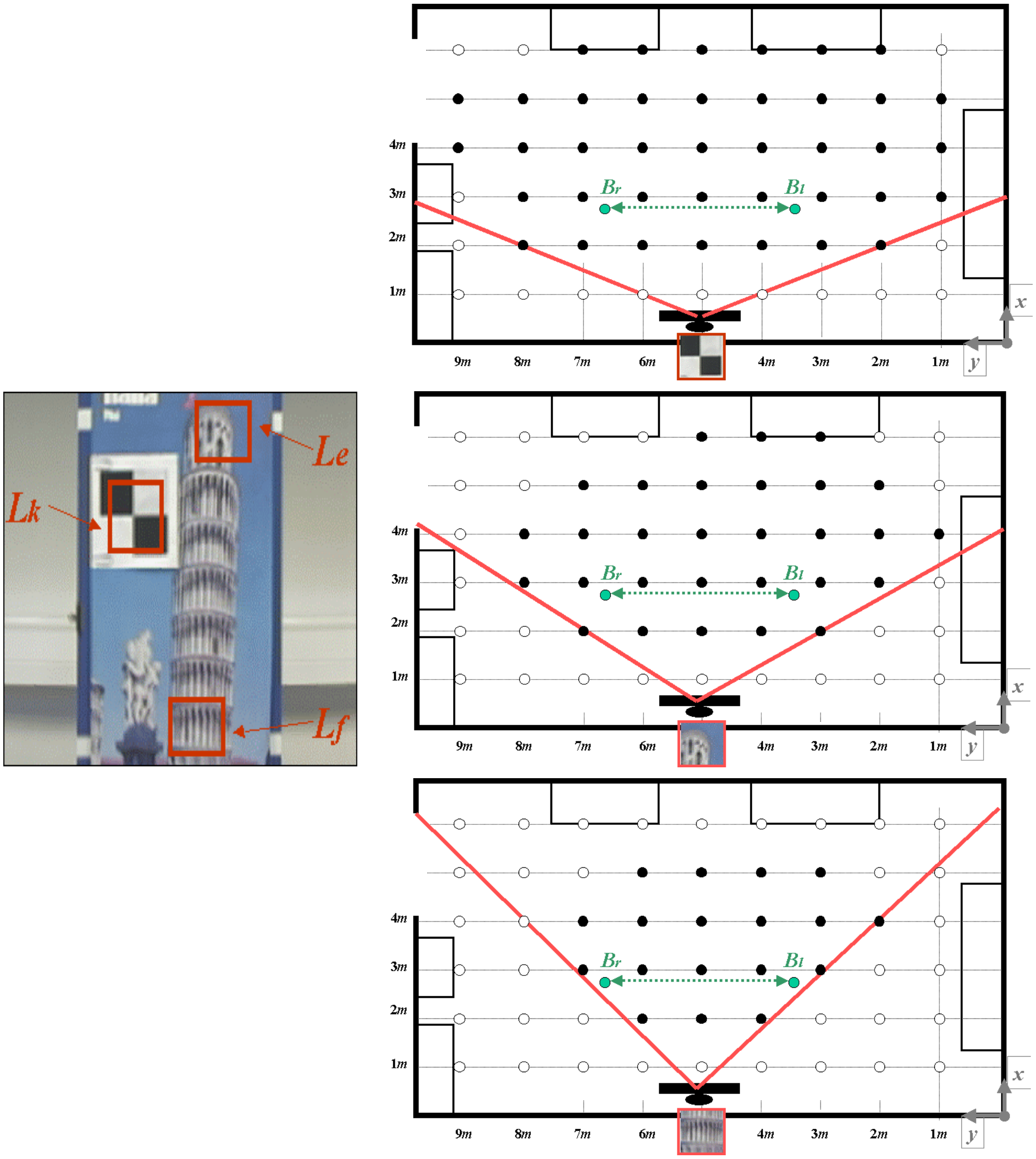


Figure 6.24: The figure left-hand side shows a panel representing three different landmarks (L_e, L_f, L_k). The figures right-hand side show workspace floor-maps which describe observation areas related to the landmarks represented in the left-hand figure.

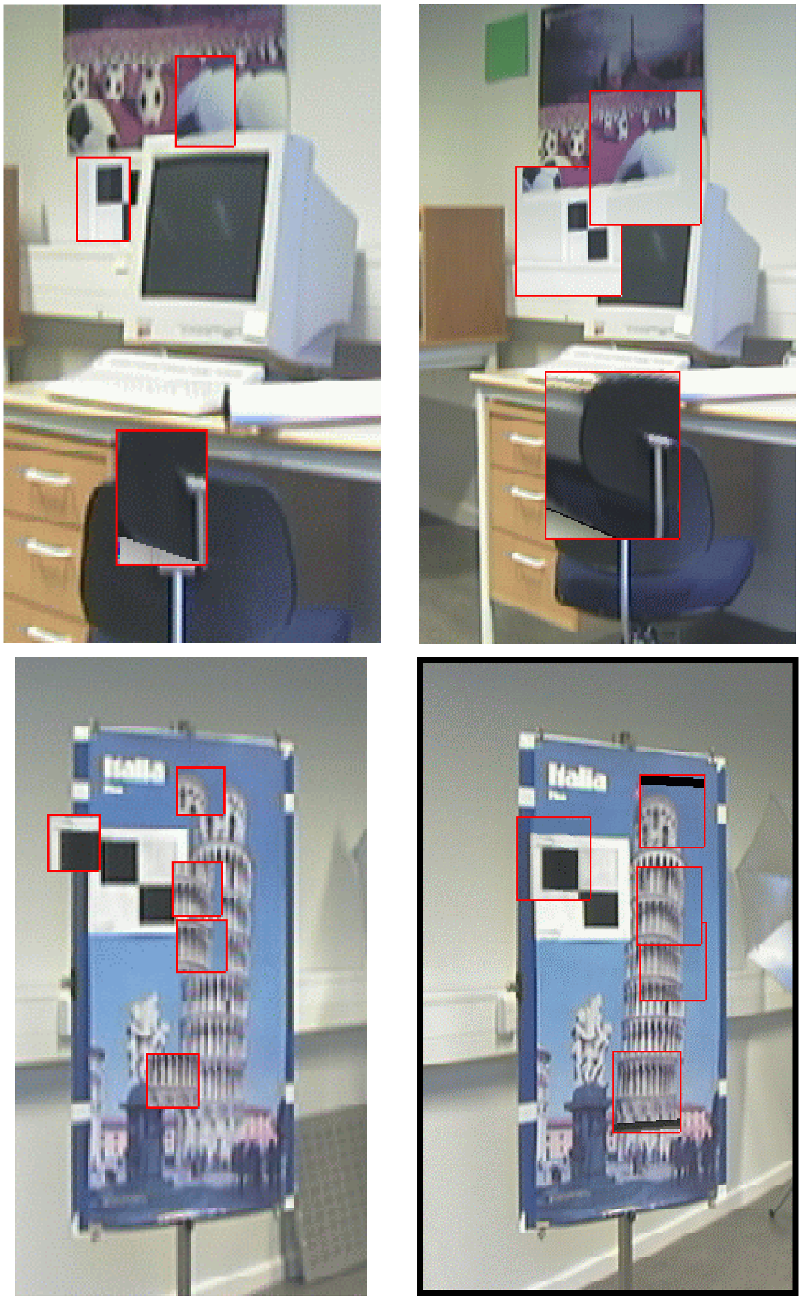


Figure 6.25: The figures show synthesized landmark visual predictions superimposed on current camera views. The left figures show different errors in predicted landmark locations. The right figures show visual predictions where landmark size is increased.

6.4.3 Summary

The results of the experimentation phase demonstrated the advantage of proposed recognition method in terms of accuracy and reliability of self-learned landmarks, and extension of observation areas.

Experiments showed that the proposed technique allows for accurate cylindrical pixel-transfer in case of considered planar or ”almost planar” landmark-surfaces, and they demonstrated a substantial improvement in recognition performance when proposed texture mapping was applied to two reference-views, compared to the case when only one reference-view was considered.

When compared to the previously proposed method related to an optimal reference-view learned by a manual training, the achieved result was not significantly worse and in some cases the result was even better.

It was interesting to notice how the system successfully matched many cases where predicted landmark views were small and distorted. This showed the power of the proposed method for acquisition and recognition of landmarks.

The match performance mainly varied according to landmark texture-structure, observation-pose, positional error, and method for visual-prediction generation. In general, the recognition performed best on landmarks having textures closer to the ideal pattern and a more accurate pose information. Natural high-contrasted patches showed lower recognition performance in term of correlation-coefficient than the ”ideal” patch. Nevertheless, they could reliably be identified in the image-plane.

Observation areas were shown to depend on landmark texture structure, positional error and landmark orientation. Observation areas represented relatively extended workspace regions, which indicated that a few recognized landmarks would be sufficient for self-localization in the new-laboratory workspace.

Chapter 7

Summary and Future Research

Undoubtedly, progress has been achieved through many years of research in mobile robotics, and it seems that nowadays service robots[1] are getting closer to becoming a reality. Mobile robotic systems, even if in cases of constrained or limited applications, have in fact been seen successfully operating in indoor and domestic settings. This has provided researchers with optimism about future perspectives, and it has partially compensated the fact that progress with autonomous mobile robots has been slower than they might have expected from the excitement and relatively rapid advances of the early days of research.

This thesis presents a new approach to autonomous robot navigation based on global localization and on the use of vision, natural landmarks, and triangulation, where the application context is a service robot. The environment is thus assumed to be indoor, semi-structured, and the robot navigates on a planar surface.

This thesis develops a self-contained system that allows a robot to accurately and continuously compute its position relative to a world coordinate system. The self-localization system is centered around the use of discriminant visual landmarks which are abundant in typical indoor environments. The thesis outlines the elements in such a system, while the particular focus is on the processes required to enable the system to automatically select, store, and subsequently recognize such landmarks from arbitrary positions, and based on the recognized landmarks estimate the absolute robot location.

In the first step, the self-localization system is tested, using visual landmarks for which the appearance of each landmark is stored as an image taken of the landmark at a known, short distance, and the world coordinate position of the landmark is measured manually and feed to the system. Subsequently, it is demonstrated experimentally that landmark appearance and position can be acquired automatically in a learning phase. Importantly it is also demonstrated that the acquired landmark positions and appearances are accurate enough to allow the self-localization system to predict landmark appearances from arbitrary positions and use this prediction to successfully recognize the landmarks. The thesis thus demonstrates the feasibility of using automatically learned visual landmarks for robot self-localization.

[1] A service robot is generally referred as a service machine that can perform various tasks in a domestic or office environment, [73].

7.1 Objectives and Proposed Idea

The research issues analyzed and developed in this thesis undertook the following issues as main objectives:

a. *Self-Localization.* The robot should be able to estimate its position automatically and accurately, i.e. with few centimeters error; and it has to keep such a positional accuracy in the medium-long range of navigation, (more than ten meters). This in turn requires the capability of recovering from localization failures. The latter is an important aspect because it represents a precondition for successful completion of most of the tasks involving service robots.

b. *Automatic Learning.* The robot should be able to build its own representation of the environment. This is partly due to the fact that installation would otherwise be too expensive, but also that the environment may change over time.

c. *Automatic Recognition.* The robot should be able to localize itself accurately also in cases when it builds its own representation of the environment. In other words, the required self-localization capability (objective "a"), should also be able to perform in case of automatically learned models, (objective "b").

The basic idea concerning the approach to be adopted for reaching the above ambitious goals, consists of proposing sensor modalities, environment models and localization algorithms, believed to have great potential, despite they are still suffering from some unreliability, which has denied them from being applied to service robots. In particular, the idea was:

(1) to explore models and technologies having a big potential in a new method for combining them;

(2) to develop and experiment with such a method by thinking consistently about the entire system concept.

To approach a solution from the system point of view was considered relevant because, as shown in this thesis, different components of a robotic system depend on each other, so that the success of the entire system depends on the performance of the various system functionalities.

It has been chosen to explore the potential of the *visual sensor*, and of environment models based on *natural landmarks*. Navigation based on landmarks represents in fact a popular alternative to having a model of the environment through which to navigate and perform self-localization. In particular, the use of visual landmarks has been proposed because of the possibility of capturing, in the form of a visual appearance rich and discriminant information. If landmarks are defined as characteristic visual structures already in the environment, (natural landmarks) then we have the possibility of fully autonomous navigation and self-localization using automatically selected landmarks.

In order to recover from localization failures and to allow for an accurate localization, (objective "a"), a global metric localization approach is proposed since this is considered the most suitable approach to accurate navigation. In particular, the localization approach is designed around a promising *triangulation* technique based on an advantageous selection of natural visual landmarks.

The system development plan consisted of three major research step, (self-localization, automatic learning, and automatic recognition) directly related to the three prefixed objectives, (points a, b, c,). The system has consequently been developed step by step, (and to decide on how to progress based on the achieved results).

It has been decided to start experimenting with landmarks which would represent planar or almost planar objects or workspace portions. The robotic system proposed for our research investigation is a mobile robotic platform which relies on a monocular vision system capable of freely panning 360 degrees. The vision unit is responsible for acquisition and recognition of visual landmarks to be used for the proposed triangulation-based localization.

7.2 Method Development

Chapter 4 presents the development of a localization method based on natural visual landmarks, triangulation, and optimal triplet selection, and it describes the experimentation of such method on the proposed robotic systems. In its first experimentation phase, the proposed system was provided with a manually trained model of the environment. The model consisted of landmark reference images and landmark position in the workspace. Based on this information, the vision system was required to: (1) identify three landmarks in the camera image-plane during navigation; (2) calculate robot position and heading by a triangulation method.

The starting point was a promising work in the state of the art providing a theoretical framework for optimal selection of landmarks triplets, (Madsen and Andersen, [97]), which could be adopted for landmark selection before triangulation takes place. However, this previous work did not demonstrate its application on a real robotic system. Consequently, the first step of the research in this thesis was to implement the proposed theory in a real robotic system. The development of this method and the experimental results are described in chapter 4. Experiments showed that accurate robot self-positioning is possible. In particular, the system was able to select optimal triplet of landmarks, recognize the associated visual patterns and precisely locate them in the image-plane, and then based on the identified location accurately estimate robot position and heading by triangulation. The system was consequently demonstrated able to recover from localization failures, i.e. errors do not accumulate. **This addressed objective "a"**.

The results of the first phase of research (self-localization) were very encouraging, and it was subsequently proposed to investigate the possibility of making such a system fully autonomous by providing the system with the capability of automatically learning the landmark model needed for the self-localization. A method mainly based on panoramic view synthesis, attention selection, and stereo reconstruction, was then proposed, developed and experimented with. The results of this research activity is described in chapter 5. Experiments showed that accurate landmark models, consisting of visual appearances of landmark and landmark pose information, could be learned by using the proposed method. The accuracy of learned information can also be estimated. **This addressed objective "b"**.

Eventually, it was necessary to verify that the automatically learned landmark positions and appearances were accurate enough to allow the self-localization system to predict landmark appearances from arbitrary positions and use this prediction to successfully recognize the landmarks. The main challenges in this case were represented by the errors associated with the learned landmarks as well as the errors affecting robot pose during its navigation. A method based on "realistic" landmark virtual-views visualization and template matching based on image-correlation was then proposed. Experiments showed that the automatically learned landmark models, could be automatically recognized by the system using the proposed method. In particular, depending on the estimated landmark positional accuracy, landmark texture structure, discrepancy in robot position between acquisition and recognition, etc., landmarks could be recognized from large extent of workspace locations. The development and experimentation of the proposed recognition method is presented in chapter 6. **This addressed objective "c"**.

The capability of recognizing automatically learned landmarks represented a very important result since with this step we had reached our last goal, (objective "c"), and so demonstrated the possibility for the proposed system to be fully autonomous. In other words, the achievement of objective "c" closes the loop, where all system functionalities have been proven through the planned sequence of consecutive steps forward.

In conclusion, the presented thesis work addresses the disadvantages related to the proposed combination of vision, landmarks and triangulation, (as presented in section 1.7), and describes a solution model able to cope and reduce the system drawbacks while allowing for an advantageous exploitation of the combination potentiality.

7.3 Contributions

The main achievements of this thesis can be outlined as follows:

- The proposed combination of techniques for panoramic-view synthesis and attention selection enables extraction of discriminant landmark views, (high-contrasted and unambiguous).
- The extracted discriminant landmark views may be used in a stereo reconstruction scheme to accurately and reliably estimate position of landmark center and surface orientation.
- Typically, the process of landmark extraction generates candidates varying in quality based on texture structure and positional accuracy. The proposed sequence of *"validity checks"* (sections 5.2 and 5.3), allows for revealing wrong estimates and for classifying learned information based on their reliability. To reveal wrong estimates as well as to classify landmark on their reliability is fundamental for improving localization performance. The latter allows for a more accurate selection of landmark optimal triplet and estimation of robot pose .
- Automatically learned landmarks can be located in the camera image-plane during robot navigation by proposed automatic recognition method, in case their positional error is a typical one. A typical error in learned landmarks is 3 cm in average (in position) and 2 degrees in average (in orientation).
- A technique from realistic virtual-view visualization was used as the core element in generating landmark view predictions. In particular, realistic virtual-view visualization

demonstrates to be a valid approach to generating visual predictions for planar and almost-planar landmarks, to be used by the proposed template matching technique. Particularly in case of almost planar landmarks results showed the proposed technique to be superior to using only one optimal reference view. The proposed technique thus allows for an advantageous use of multiple landmark-views accumulated during the learning phase.

- The template matching algorithm for landmark recognition based on a normalized cross-correlation shows to be suitable to recover typical landmark positional errors. The size of the search windows plays an important role.

- The proposed method for robot self-localization based on triangulation and optimal triplet selection may provide a very accurate robot positioning. In particular, this technique allows for global metric localization and recover from accumulated errors and localization failures. Experimental runs showed a typical error of 3 cm. in robot position and 3 degrees in robot heading.

- The proposed optimal landmark-triplet selection is fundamental. In fact, it allows for extended period of autonomous navigation (the uncertainty is maintained low) based on an advantageous exploitation of the available information associated to landmarks, (landmark position and orientation, distribution of associated uncertainty, observation areas).

The main contributions of this thesis are considered to be on two levels: more general (point 1), and more specifical (points 2, 3, 4). In particular:

1) The combination of vision and natural landmarks for accurate global localization and an approach description for the entire (navigation) system main components. The State of the Art revealed that this was something not tried before.

2) The use of realistic virtual-view visualization technique for generating landmark visual prediction. The novel aspect is the proposed use of these techniques for the purpose of image matching, (rather than just visualization).

3) The use of a global localization technique based on triangulation and optimal triplet selection, for the purpose of accurate robot self-positioning.

4) The strategy for automatic learning of a landmark-based environment model, by the extraction of discriminant visual appearances and accurate large-baseline stereo-based position estimates.

7.4 Future Research

Some aspects of proposed system were not addressed in the thesis and still represent open issues which could be of interest for further developments. Among them: a further system integration, a more "adaptive" estimation of the uncertainty, and a wider range of landmark "shapes".

- **System Integration**. A future work would see the system furtherly integrating the tested self-localization system based on triangulation, with the landmark acquisition and recognition sub-systems, which have all been demonstrated in this thesis. This would in fact allow for refining proposed algorithms in the different system modules, while better "tuning" the involved parameters. This will also lead to get a deeper understanding about how the system parameters should be set according to the type and dimension of the environment surrounding the robot. Hence, to investigate extension to large environments of proposed localization approach. The integration of proposed learning with navigation following the proposed learning navigation strategy (section 5.4) also represent a future work.

- **Adaptive Uncertainty**. Additional improvements could be achieved by an "adaptive" estimation of landmark positional uncertainty, which could be calculated based on the current error expected in robot pose at the time of the learning. In performed experiments such uncertainty was set as constant on the basis of a typical error measured during self-localization trials. Such a choice was supported by the hypothesis that the robot could arbitrarily decide to stop during its navigation to learn new landmarks, so that we wanted to demonstrate that the robot was able to learn useful landmarks in such situation. However, when the current situation is such that the system initializes robot pose anew, (e.g. after having estimated a large uncertainty or at initialization), it would be useful to re-set expected uncertainty in robot pose to a smaller value, (typical after pose initialization), before a new learning phase takes place. This would lead to higher fidelity in landmark associated uncertainty.

  The estimated uncertainty in landmark and robot pose should also be used to dynamically size and shape correlation search-windows in the camera image-plane, before template matching. This aspect was not developed in the current system implementation because experiments were focused on testing the limits of proposed approach. However, this aspect is very critical towards system performance. The type and shape of search-window proposed by Davison, [41], could be an example to follow.

  The method for optimal triplet selection also needs further research in order to adaptively include landmark positional uncertainty. In particular, in order to estimate how such uncertainty should be evaluated when selecting an optimal landmark triplet in relation to current robot position and its estimated accuracy. It should also be investigated a suitable way of combining expected uncertainty in landmark pose and image-plane location, (see subsection 4.2.2), when selecting the optimal triplet.

- **Landmark Shape**. The proposed "medium size" landmark patch used in typical indoor environment usually represent "large" workspace portions. This requires landmarks to represent planar or "almost-planar" environment surfaces. In fact, different landmark views need to show similar texture-structure to be matched, which requires landmarks to be either "far away", observed with a short baseline configuration, or lying on a planar or an "almost-planar" surface. Despite a typical indoor environment has plenty of planar surfaces, this aspect represents a limitation for the proposed method since lots of potentially suitable objects can not be used, which may become critical in case of environments possessing large planar surfaces, such as walls, with little or no texture.

  An interesting aspect which could be investigated in future research is consequently related to the development of a learning method able to cope with landmarks having shapes more complex than planar or "almost planar". It is in fact believed that a further investigation on the application of techniques from realistic virtual-view visualization to objects with a more complex shape, could bring to dramatic improvements in term of recognition performance. In particular, from a first circumscribed analysis it appeared that landmarks which shape could be approximated by few planar surfaces, would still be suitable to be used in a recognition scheme based on image correlation. In this case, landmarks could, for example, be defined as a collection of few planar surfaces and their relative spatial configuration, which could then be matched in an elastic template matching scheme, such as in Balkenius [9].

  In case of landmark shapes which would lead to models more complex than just a collection of few surfaces, the use of techniques from realistic virtual-view visualization could still be successful. Nevertheless, in these case the research should perhaps consider matching techniques different than image correlation. This would better cope with the problem of approximated landmark view predictions. The use of feature-based image matching techniques (see for example Matas et al. [101], Zingaretti and Carbonaro [152]), could then become more appropriate. Nevertheless, other image matching approaches should also be considered and evaluated. Among them, appearance based-methods (De Verdiere and Crowley [42], Zingaretti et al. [154]), and analysis of grey-level invariants (Schmid and Mohr [124]).

The presented thesis seems to open up to a new potential way of providing a mobile robot with accurate and reliable global localization capability, based on the potential of the combined use of vision, natural landmarks, triangulation and optimal landmark selection. With this view, future investigations could lead to significant achievements in the field, and towards the development of applications of such a method in commercial robots operating in indoor settings.

Appendix A

Main Approaches in Realistic Virtual View Synthesis

This appendix provides a brief overview of main reference approaches in the field of realistic virtual view synthesis. The appendix is intended for a reader interested in knowing more than what described in subsection 6.2 about image- and model-based rendering approaches, but not as an exhaustive survey of the research field.

The synthesis of realistic virtual views has received increasing attention in the last decade, mainly due the increased popularity of virtual reality and the spread of its applications. Hence, to the demand of increasing realism into generated sceneries while simplifying the modeling process.

A solution to the problem of providing visual realism to computer generated images is searched into the possibility of re-creating naturally occurring physical phenomena by real world observations. The "road" mainly investigated has been to capture the occurring phenomena through photographs, hence, directly or "indirectly" transfer them into novel generated views.

The direct way refers to warping algorithm which do not take into consideration any geometric information associated to the observed scene. These approaches usually require a dense set of reference-views. The "indirect" way instead refers to the case when reference image-views are supported by associated knowledge, (pixel correspondences, depth maps), or 3D models.

The achievement of such ambitious goal, a realistic synthesis, has mainly be attempted through two different rendering approaches: model-based and image-based.

The model-based approach represents the traditional way to generate virtual views of an object or scene. This approach is usually referred as *Model-Based Rendering* because it relies on a geometrical 3D model of the object or scene wished to be rendered. In this context the research is focused on improving model fidelity by using image-based modeling. In particular: geometric model extraction and representations for rendering purposes, object-texture extraction and mapping to geometric models, illumination effects recovery and rendering.

At a high level a model-based rendering approach involves three processes. First, an event or a scene must be recorded, then, a 3D model of the environment has to be extracted using computer vision techniques, and at the end, the obtained 3D model is rendered from the view of a virtual camera.

The image-based approach represents instead an alternative to model-based rendering, and a competing means of creating virtual views, primarily relying on real images taken as reference in place of a geometric 3D model. This approach is then referred as *Image-Based Rendering.* In order to produce novel views, reference images are usually interpolated or re-projected from source to target image.

This rendering approach is less generic than model-based rendering since utilized techniques often depend on the applications, thus, type of environment and required rendered field of view. The common characteristic is a rendering time independent from scene complexity and no need in principle for reconstruction of geometric models.

Image-based rendering techniques often require an additional knowledge to input reference images, such as image-correspondences, depth information, epipolar relations, etc. This additional knowledge often is extracted from same input images or it is provided a priori.

Image-based rendering is usually applied to static environments whereas model-based rendering is often proposed for dynamic scene visualization. Authors have also proposed both the two approaches for the same application context, (Blanc, Livatino and Mohr [14], [16]).

A growing interest, which could also be considered as an "evolution" of image and model based rendering is towards hybrid methods, so that in the last years authors have presented methods which lie in-between image- and model-based rendering. A successful example is represented by the work of Debevec, Taylor and Malik [43], which proposes generation of novel views based on a reconstructed geometric model, where textures in novel views are mapped view-dependently.

Survey papers have also started with classifying works in realistic virtual view synthesis as a "continuum" of representations depending on the amount of geometry, (Shum and Kang [130]), and amount of geometry related to number of reference-images, (Buehler et al. [18]).

A.1 Image-Based Rendering

As above mentioned, the image-based approach relies on real images taken as reference in place of a geometric 3D model. However, input reference images often do not suffice for the purpose of rendering novel views, so that many of the proposed systems require additional knowledge, such as image-correspondences, depth information, epipolar relations, etc.

The advantage represented by avoiding model reconstruction may also mean software rendering and no exploitation of progress in graphic hardware, (unless the system has been designed for the exploitation of some graphic functions, e.g. projective texture mapping).

Image-based rendering methods often mentioned in current literature, were developed in the half of the nineties, (Chen [29], Levoy and Hanrhan [85], Gortler et al. [59], McMillan and Bishop [103], Seitz and Dyer [125], Chen and Williams [28], Shashua and Werman [128], Chang and Zakhor [25], Leveau and Faugeras [84], Regan and Pose [116]). Research is still very active in the field and new techniques have also been proposed for investigation. For example, the new projection model based on the *two-slit camera*, (Granum et al. [61]).

In this section summaries of representative works in Image-Based Rendering are presented. The authors name at the top of each summary identifies presented approach together with a "pioneer" reference publication. Figure A.1 represents typical computational steps involved in Image-Based Rendering.

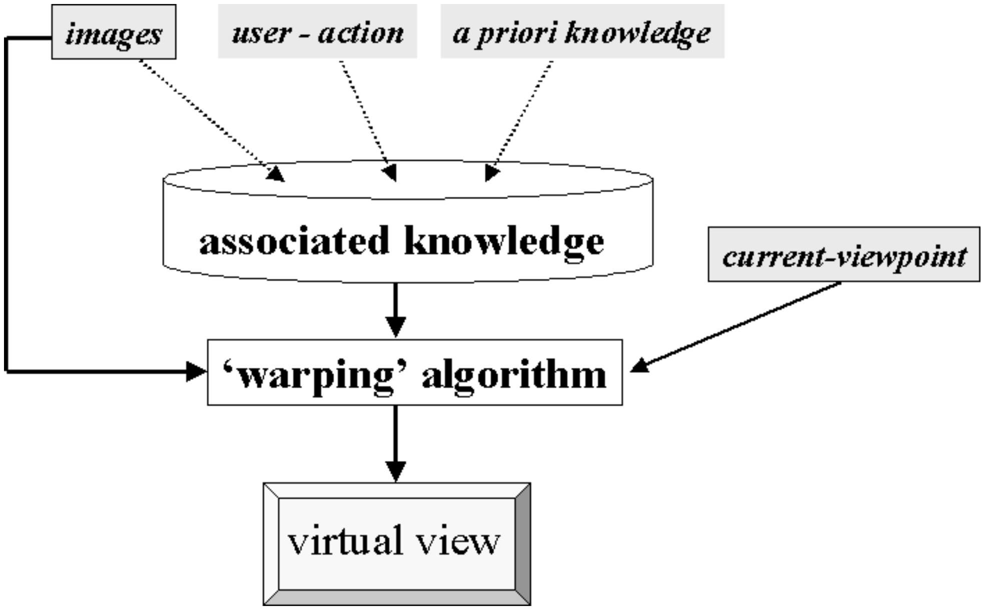


Figure A.1: Image-based rendering: typical computational-steps.

Chen [29]

S. E. Chen proposes an image-based rendering system called *QuickTimeVR* developed by Apple Computer.

The paper presents a way for a computer to systematically deal with movies. It includes an algorithm for storing moving pictures and play them back as fast as possible without using any extra hardware. The system includes a "panoramic movie" technology which enables users to explore spaces, and an "object movie" technology which enables users to examine objects interactively. This system is used on the world wide web to display 3D objects from various viewpoints.

The scene is represented by a set of cylindrical images created at key locations. Based on these images the system is able to synthesize new planar views in response to user input by warping one of these cylindrical images. The user is so able to navigate "discretely" from location to location, and while at each location continuously change the viewing direction. Translation of viewing position can instead only be approximated by selecting reference cylindrical images closest in viewpoint to current viewing position. The above is achieved at interactive rates (greater than 20 frames per second).

The speed of the processor largely determines the quality of visualized movies. If the system can not process all frames in the movie, QuickTimeVR drops some frames. The system can be applied to exchange video on the Internet, virtual navigation of real environment, (useful for architecture planning, museum tours), etc.

Among the advantages: the method represents a practical way of exchanging video on the Internet, virtual navigation of real environment, allows for immersive navigation of visual environment. No need for considering the viewing angle when selecting a reference images, (references are cylindrical), no need for specialized hardware, high-quality images, distortion corrector, multimedia possibilities.

Among the disadvantages: visualized scenes must be static, only playback visualization, several photographs are required and properly registered.

Class of Approaches: no-geometry rendering [130], light-fields [54], mosaicking [78].

Levoy-Hanrahan [85]

This paper describes *Light Field Rendering*, a simple and robust method for generating new views from arbitrary camera positions without depth information or feature matching, simply by combining and re-sampling the available images.

The major idea behind the technique is a representation of the light field, the radiance, as a function of position and direction, in region free of occluders. In these regions the "light field" is a 4D parameterization of viewing position and direction. An image is a two dimensional slice of the 4D light field. Creating a light field from a set of images corresponds to inserting each 2D slice into a 4D light field representation. Similarly, generating new views corresponds to extracting and re-sampling a slice. Once a light field has been created new views may be constructed in real time by extracting slices in appropriate directions. The desired ray can be looked up in the light field database of rays using the 4D parameterization of viewing position and direction.

Image generation using light fields is inherently a database querying process, much like the movie map image-based approach of Chen [29]. The interpolation scheme used by the authors approximates the re-sampling process by simply interpolating the 4D function from the nearest samples. The authors have investigated the effect of using nearest neighbor, a bilinear interpolation, and full 4D quadrilinear interpolation.

Since the success of the method depends on having a high sample rate, the authors describe a compression system that is able to compress the generated light fields by more than a factor of 100:1 with very little loss of fidelity. A vector quantization scheme is used to reduce the amount of data used in light field rendering, yet achieving random access and selective decoding. The authors have also addressed the issues of anti-aliasing during creation, and re-sampling during slice extraction. In particular, to reduce aliasing effect, the light field is pre-filtered before rendering.

Among the advantages: real-time display of new views by extracting slices in appropriate directions, high freedom in the range of possible views, no model information such as depth-values or image-correspondences is needed to extract the image values, image generation involves only re-sampling (a simple linear process), simple compression schemes can be applied (because of the 3D structure of the light field), re-sampling process simpler than depth or correspondence -based, image-based rendering approaches.

Among the disadvantages: large amount of data that may be required, (but possible high compression by the proposed method), long time for image acquisition (reference images are acquired by scanning a camera along a plane using a motion platform), the flow of light is completely characterizes only through unobstructed space in a static scene with fixed illumination, sampling density must be high (to avoid excessive blurriness).

Class of Approaches: no-geometry rendering [130], light-fields [54], interpolation from dense matching [78].

Gortler-Grzeszczuk-Szeliski-Choen [59]

This paper discusses *The Lumigraph*, a new computational method for capturing the complete appearance of both synthetic and real world objects and scenes, representing this information, and then using this representation to render images of the object from new camera positions.

The Lumigraph as in the case of light-field rendering is a ray-database query algorithm. The lumigraph uses a 4D parameterization of viewing position and direction, (a 4D parameterization of rays passing through a pair of planes with fixed orientation). The lumigraph, unlike the light field, considers the geometry of the underlying models when reconstructing desired views. The geometric information is used to control the blending of the images. The lumigraph, as well as the light field, is defined as data intensive rendering process. However, the lumigraph can tolerate a lower sampling density since the available geometric information.

Among the advantages: fast scene rendering, arbitrary camera poses are used to construct the database of visible rays, high freedom in the range of possible views.

Among the disadvantages: the preparation of the database requires considerable pre-processing, large amount of data that may be required, the flow of light is completely characterizes only through unobstructed space in a static scene with fixed illumination, sampling density must be high.

Class of Approaches: no-geometry rendering [130], light-fields [54], interpolation from dense matching [78].

McMillan-Bishop [103]

L. McMillan and G. Bishop propose *Plenoptic Modeling* as a consistent framework for the evaluation of image-based rendering systems. The authors give a concise problem definition and propose an image-based rendering system in light of the Plenoptic framework.

The paper introduces the use of the 5D plenoptic function, $P_5(V_x, V_y, V_z, \theta, \phi)$, defined as the intensity of light rays passing through the camera center at every space locations (V_x, V_y, V_z) at every possible angle (θ, ϕ). The original 7D plenoptic function was presented by Adelson and Bergen [1]. The simplest plenoptic function is a 2D panorama, cylindrical or spherical, when the viewpoint is fixed.

Within the proposed plenoptic modeling the goal of image-based rendering is to generate a continuous representation of the Plenoptic function. The authors claim that all image-based rendering techniques can in fact be casted as attempts to reconstruct the Plenoptic function from a sample set of that function. They believe there are significant insights to be gained from this characterization, so they propose their system in light of this Plenoptic framework.

The samples used are cylindrical panoramas. The "angular disparity" of each pixel in stereo pairs of cylindrical panoramas is computed and used for generating new plenoptic function samples. The authors also introduces a geometric invariant for cylindrical projections that is equivalent to the epipolar constraint defined for planar projections. The original samples, cylindrical panoramic images, can so be used to reconstruct new virtual views from arbitrary locations. The reconstructed views are also capable of describing perspective effects and occlusions. In particular, the authors introduce a novel visible surface algorithm which guarantees back-to-front ordering.

Among the advantages: real-time display of visually rich environments (both indoor and outdoor) is possible without the need for special graphic hardware, the method allows for acquisition and exploitation of compact sample images, realistic visualization of complex sceneries where perspective effects and occlusion are correctly modeled, real-time display and the use of commonly available equipment.

Among the disadvantages: visualized scenes must be static and with fixed lighting conditions, reference images should be acquired close to each other, reconstructed views should be generated close to sample images.

Class of Approaches: no-geometry rendering [130], light-fields [54], mosaicking [78], geometrically-valid pixel reprojection [78].

Seitz-Dyer [125] [126]

S.M. Seitz and C.R. Dyer propose *View Morphing*, a way to generate new views of a scene from two basis views. It can be applied to both calibrated and uncalibrated images. At minimum, two basis views and their fundamental matrix are needed.

A scan-line algorithm for making image interpolation is presented that require only four user provided feature correspondences to produce valid orthographic views. The paper describes a simple image rectification procedure which guarantees that interpolation does in fact produce valid views, under generic assumptions about visibility and projection process.

The proposed technique uses basic principles of projective geometry, and introduces an extension to image morphing that correctly handles 3D projective camera and scene transformations. The authors propose to exploit monotonicity along epipolar lines to compose physically valid intermediate views without the need for full correspondence information. Under the assumption of monotonicity, it is shown that the problem is theoretically well-posed.

This result is significant in light of the fact that is not possible to fully recover the structure of the scene due to the aperture problem [1]. Moreover, they demonstrate that for a particular range of views, the problem of view synthesis is in fact well-posed and does not require a full correspondence, that is, images interpolation is a physically valid mechanism for view interpolation. Views can consequently be generated by linear interpolation of the basis images, (if the basis images are first rectified).

Among the advantages: the method represents a practical and simple way of generating new views of a scene (under monotonicity assumptions), view synthesis does not suffer from the aperture problem, the technique may be applied to photographs as well as rendered scene, ability to synthesize changes both in viewpoint and image structure, interesting 3D effects via simple image transitions, applicable to both calibrated and uncalibrated images, suitable for application in entertainment industry and for limited bandwidth teleconferencing.

Among the disadvantages: the method requires multiple image re-sampling (loss of quality), local blurring when monotonicity assumption is violated, artifacts arising from errors in correspondence, it is only suitable for static scenes, the method needs four user provided feature correspondences, visualized regions need to be free of occluders.

Class of Approaches: implicit-geometry rendering [130], volumetric reconstruction [54], geometric-valid pixel reprojection [78].

Chen-Williams [28]

This paper presents *View Interpolation* an image interpolation approach to synthesize 3D scenes, where input images are a structured set of views of a 3D object or scene.

In order to reconstruct desired views several reference images are used along with image correspondence information. The view synthesis is based on linear interpolation of corresponding image points using range data to obtain correspondences, (as in view-morphing [125]).

Intermediate frames are used to approximate intermediate 3D transformations of the object or scene. The authors have investigated smooth interpolation between images by modeling the motion of pixels (i.e. optical flow) as one moves from one camera position to another. They have investigated special situations in which interpolation produces valid perspective views. They conclude that interpolated images do not in general correspond to exact perspective views. They point out and suggest solution for determining the visible surfaces. Like image morphing, View Interpolation uses photometric information as well as local derivative information in its reconstruction process.

Among the advantages: the proposed method can be performed at interactive rates, suitable for virtual holograms, walk-trough in virtual environments, incremental rendering, motion blur acceleration, and soft shadows cast (by area light sources) acceleration, the approach works well when generated views share a common gaze direction and the synthesized viewpoints are within 90 degrees of this gaze angle.

[1] The Aperture problem arises due to uniformly colored surfaces in the scene. In the absence of strong lighting effects, a uniform surface in the scene appears nearly uniform in projection. It is then impossible to determine correspondences within these regions.

Among the disadvantages: problems in the generated images for points which are not mutually visible on both reference images (difficult to establish the flow field information), view approximation when the change in viewing position is not slight, static scene, problems may arise when the generated views do not share a common gaze direction, and when the synthesized view-points do not stay within 90 degrees of the gaze angle.

Class of Approaches: implicit-geometry rendering [130], volumetric reconstruction [54], interpolation from dense matching [78].

Shashua-Werman [128]

This paper based on the existence of certain trilinear functions of three views, (with a corresponding tensor of 27 intrinsic coefficients), [127], derives connections between the trilinear function invariants across three views and intrinsic structures and invariants of 3D space.

The result shows that the tensor of coefficients determined by three views replaces entirely the role of the fundamental matrix (and associated intrinsic structures of two views) in 3D tasks. In other words, the projective structure of the scene follows directly from the tensor without the need to recover any intrinsic structure associated with two views.

In addition the tensor encompass 2-view structures in the sense that the fundamental matrix is readily expressed as a solution of a linear system determined by the tensor, the rotational component of camera motion is expressible in closed form by the tensor, and a variety of means exist for recovering the epipoles from the tensor.

The major result is that exists a decomposition of the tensor into three matrices that corresponds to three intrinsic homography matrices of the three distinct planes. The planes are associated with the camera coordinate frame of the third view and provide a reference basis for reconstruction of invariants. This provides a geometric intrinsic structure of three views.

The authors claim that the tensor offers a host of a new algorithms for recovering 3D information from 2D views, cuts through the epipolar geometry, makes room for statistics, and generally exploits the information available from measurement across views in a more efficient manner than any technique based on 2-view geometry.

Among the advantages: new algorithms for recovering 3D information from 2D views, an order of magnitude improvement compared to conventional techniques that rely on epipolar geometry (when synthesizing novel views from a pair of model views), applications in virtual reality, 3D television, recognition, fast rendering, 2-views structures (fundamental matrix, epipoles) are recoverable (linearly) from a tensor.

Among the disadvantages: static scene, some fiducial points are needed.

Class of Approaches: implicit-geometry rendering [130].

Avidan-Shashua [8]

This paper proposes a method where views are reconstructed directly without first estimate the depth, by exploiting certain invariants in the geometry of the problem.

Input consists of 3 images from which it is possible to compute a trilinear tensor who will provide a correct way to generate virtual views of the observed object. In particular, the trilinear tensor is computed from the point correspondences between reference images. In case of only two images, one of the images is replicated and regarded as third image. If the camera intrinsic parameters are known, then a new trilinear tensor can be computed from the known pose change with respect to the third camera location. The new view can subsequently be generated using the point correspondences from the first two images and the new trilinear tensor.

The authors claim that the trilinear tensor gives user wider perspective transformation possibilities than other methods in literature. Texture is achieved by an interpolation of reference images. A realistic effect is achievable with this technique, however, image rendering might not be real-time because of the dense matching and tensor computation.

Among the advantages: realistic effect, the use of tensor (recovering 3D information from 2D views, no epipolar geometry, etc), efficient synthesis of novel views and wide visualization range, applications in virtual reality, 3D television, recognition, fast rendering, 2-views structures (fundamental matrix, epipoles) are recoverable (linearly) from a tensor.

Among the disadvantages: this approach does not correctly reconstruct points that become occluded.

Class of Approaches: implicit-geometry rendering [130], points transfer [54].

Laveau-Faugeras [84]

The authors propose a system where views are reconstructed directly without first estimate the depth. Under the assumption that a complete pixel-wise correspondence is available, it is possible to predict a broad range of views. The use of epipolar geometries between images restricts the image flow field in such a way that it can be parameterized by a single disparity value and a fundamental matrix which represents the epipolar relationship. The authors also provide a two-dimensional ray-tracing-like solution to the visibility problem which does not require an underlying geometry description. Their method does, however, require establishing correspondence for each image point along the ray's path.

Class of Approaches: implicit-geometry rendering [130], points transfer [54].

Chang-Zakhor [25], [26]

This paper presents a method to generate arbitrary views of three dimensional scene by means of an intensity-depth representation.

By using an uncalibrated camera which scans a stationary scene under approximately known camera trajectories, and then by transforming points on camera image planes onto the plane of the virtual view, the proposed system derives dense depth-maps at several preselected viewpoints.

The authors propose an adaptive matching algorithm which assigns various confident levels at various regions. Once the depth maps are computed at preselected viewpoints, the intensity and the depth at these locations are estimated using a stereo algorithm and used to reconstruct arbitrary views of the 3D scene.

Among the advantages: fast and flexible image acquisition (hand held cam-corder, uncalibrated cameras, unknown camera position), well estimate depths, image-quality good for the most part, well-reconstructed horizontal edges, few errors concerning occluded regions.

Among the disadvantages: artifacts due to specularities of the surface, image matching performed poorly for background regions which are seen through holes of foreground regions, static scene (stationary 3D objects), only horizontal motion.

Class of Approaches: geometric-valid pixel reprojection [78].

Rousso-Peleg-Finci [118]

This paper concerns with stitching together images from adjacent viewpoints in order to generate a realistic panoramic virtual view of an observed environment. The authors propose an algorithm based on the method proposed in [114], to solve the main problem of panoramic mosaicing which is related to the forward camera motion (e.g. zooming). Pictures are segmented in vertical strips which are aligned by a "stretching" technique. In this way distortions appear greatly reduced.

CohenOr [32]

This paper presents a way to exploit projective texture-mapping to render adjacent views of reference images. The authors called these views *Extrapolated Views.* The aim was to improve time-performance of a walk-through in remote virtual environment.

Hirose [70]

This paper proposes the use of a camera with position sensors in order to make a interactive walk-through, based on pre-recorded sequence of images which are stored in a database. The use of image interpolation greatly reduces the required number of pre-recorded images while the known image-position allows to the system to recover reference images of interest from the database.

Regan-Pose [116]

This paper describes a hybrid system in which plenoptic samples are generated on the fly by a geometric-based rendering system at available rendering rates, while interactive rendering is provided by the image-based subsystem. At any instant, a user interacts with a single plenoptic sample. The authors also discuss local reconstruction approximations due to the changes in the viewing position. Local reconstruction approximations amount to treating the objects in the scene as being placed at infinity, resulting a loss of kinetic depth effects.

A.2 Model-Based Rendering

Model-based Rendering usually recovers the geometry of the real scene and then render it from desired virtual view points. Methods for the automatic construction of 3D models have found applications in many field, including Mobile Robotics, Virtual Reality and Entertainment. These methods generally fall into two categories, active and passive methods.

Active methods often require laser technology and structured lights or video, which might result in very expensive equipments. However, new technologies have extended the range of possible applications, (Levoy et al. [86], Hogg et al. [71], Fisher et al. [53]), and new algorithms have improved the ability to cope with problems inherent to laser scanning, (Castellani, Livatino and Fisher [24], [23], Stulp [135], Davis et al. [38]).

Passive methods usually concerns the task of generating a 3D model given multiple 2D photographs of a scene. In general they do not require a very expensive equipment, but quite often a specialized set-up, (e.g. Kanade et al. [77], Fuch et al. [55], Tseng and Anastassious [145]). Passive methods are commonly employed by Model-Based Rendering techniques.

Some of the research contributions in the field have proposed fully working systems for specific applications, some other have instead mostly focused on one or some of the involved aspects but provided a general application context. For example, Moezzi et al. [105], [106], propose an entire specific system for image-acquisition, model-construction and play-back interactive rendering, while Ofek et al. [110], mostly focus on extraction of textures from a generic video sequence for high-fidelity model-based texture mapping.

There can well be different ways of generating 3D models from photographs, from simple 3D silhouette models dynamically cut-out and texture-mapped from video-sequences, (Livatino and Hogg, [91]), to polyhedral visual hulls generated by multiple-view silhouettes [2] and estimated stereo disparities, (Li, Schirmacher and Seidel, [87]). A survey of image-based volumetric scene reconstruction can be found in the works of Slabaugh et al. [131], Dyer [49], and Forsyth and Ponce [54].

In this section summaries of some representative works in Model-Based Rendering are presented. The authors name at the top of each summary identifies presented approach together with a "pioneer" reference publication.

Figure A.2 represents typical computational steps in presented works.

[2] The visual hull represents a conservative shell that envelops the true geometry of the objects, consisting in the shape obtained from silhouette image data. The visual hull techniques require that foreground objects in the input images can be segmented from the background.

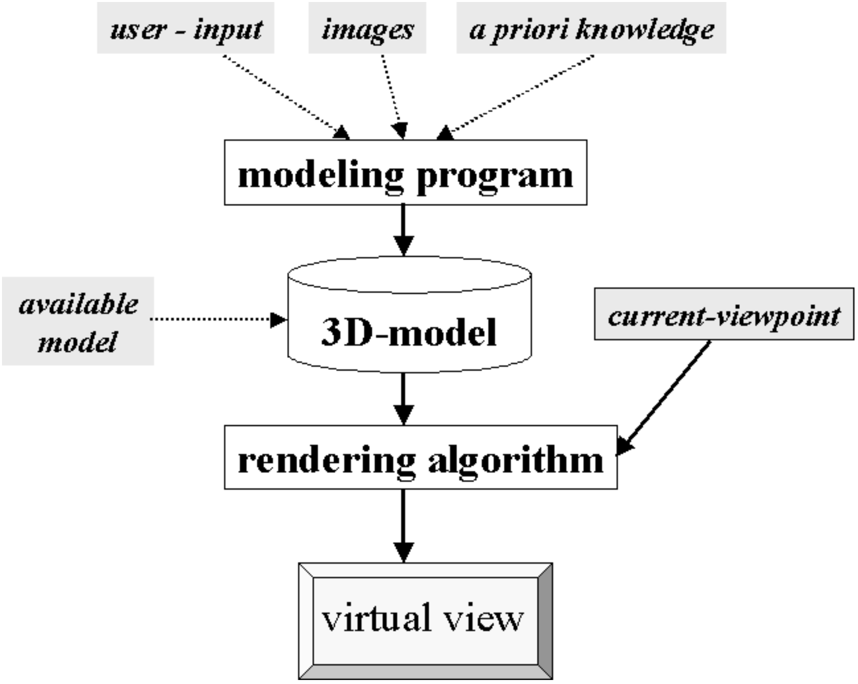


Figure A.2: Model-based rendering: typical computational-steps.

Kanade-Narayanan-Rander [77]

Kanade et al. coin the term *Virtualized Reality* to characterize a system that is able to capture dynamic scenes and render them from different virtual viewpoints. This in order to immerse viewers in a virtual reconstruction of real-world events.

A visual event such as an actor motion is captured using many cameras (from six cameras to many more) placed all around a hemispherical dome 5 meters in diameter that cover the action from all sides. Consequently, several real-images of the scene are acquired.

The 3D structure of the event, aligned with the pixels of the image, is computed for a few selected directions using a stereo technique. In particular, depth information are recovered by stereo matching and the combination of depth/color data converted into a triangle mesh model on graphics rendering engine.

The authors use a multi baseline stereo algorithm to compute time-varying 2.5-D depth maps representing the scene geometry. Stereo-based depths are aligned with the pixels of their corresponding images.

Based on the viewer's position, the depth map from the closest camera is used to render the scene. Triangulation and texture mapping enable the placement of a "soft-camera" to reconstruct the event from any new viewpoint. Virtualized reality allows a viewer to move freely in the scene, independently from the angles used to record the scene.

Among the advantages: the system provides 3D structure of events for virtual reality applications, safe training, user guided visualization of events (for entertainment).

Among the disadvantages: the system does not allow for an on-line processing, (but only play-back) due to the high computational cost.

Class of Approaches: volumetric reconstruction [54].

Fuchs-Bishop-Arthur-McMillan-Bajcsy-Lee-Farid-Kanade [55]

This paper proposes the use of image data acquired by many stationary cameras installed around a small environment such as a conference room and the use of stereo methods to compute time-varying 2.5-D depth maps representing the scene geometry.

The authors propose to reconstruct the real world scene from a large amount of fixed cameras by applying a correlation based depth from stereo. Wide baseline stereo is used to extract depths maps, which are updated, maintained, and then combined to create a virtual scene from viewer current position and orientation. The data can be acquired in a remote site while the viewer position and orientation is local.

In shown results a depth image of a human subject is calculated from 11 closely spaced video camera positions. The user is wearing a head-mounted display and walks around the 3D data that has been inserted into a 3D model of a simple room.

Among the advantages: the system is suitable for teleconferencing applications, provide 3D structure of events for virtual reality applications, safe training, user guided visualization of events (for entertainment).

Among the disadvantages: only play-back is allowed, (the speed of the stereo algorithm was much limited by the poor machine performance).

Tseng-Anastassiou [145]

This paper proposes the use of images captured simultaneously by a set of equi-distant cameras with parallel axis, in vertical and horizontal lineups and the use of stereo methods to compute time-varying 2.5-D depth maps representing the scene's geometry. Virtual images are generated by interpolating real views scanline-by-scanline based on disparity information.

Within the MPEG standardization the transmission of a stereoscopic (left and right views) sequence is possible by utilizing the proposed high profile double layer structure of temporal scalable coding. The left stereo sequence is coded on the lower layer and provides the basic non-stereoscopic signal. The right stereo bitstream is then transmitted on the enhancement layer and when combined with the left view results in the full stereoscopic video. After decoding the two extreme views, an "intelligent" scheme is proposed to interpolate the intermediate views.

Among the advantages: MPEG can be applicable to two sequences of stereoscopic signals through the use of spatial and temporal scalability extensions, easy and direct implementation convenience, graceful stereo image degradation, and high SNR reconstructions, compression, and improvement in image reconstruction.

Moezzi-Tai-Gerard [106]

This paper proposes to recreate the original dynamic scene in 3D, the system allows photo-realistic interactive playback from arbitrary viewpoints using video streams of a given scene from multiple perspectives.

The idea is to capture multiple images of an object and then construct a 3D textured model of an object and use view-dependent texture mapping for rendering (from any view angle). The authors use 17 cameras surrounding a stage area to record various performances. The 3D model is extracted by an accurate recovery of the 3D shapes of dynamic or foreground objects by means of a volume occupancy method.

This work is based on Moezzi et al. [105] who construct a visual hull, i.e. a conservative shell that envelops the true geometry of the objects, consisting in the shape obtained from silhouette image data. The visual hull is constructed using voxels in a off-line processing system. The shape of the visual hull can be determined from object silhouettes in multiple images taken from different viewpoints, (the silhouette information is obtained by background subtraction).

In order to render the obtained model from a virtual camera point of view, a true 3D model is created with fine polygons, each separately colored. There is no need for texture rendering support and the viewing position plays no role in the modeling process.

The proposed approach can use standard object formats such as VRML delivered through the Internet and viewed with VRML browsers. Hence, the approach is suitable to the client-server scenario because real views do not need to be transferred to the client.

Among the advantages: accurate 3D model reconstruction, no texture rendering support needed, use of VRML format for browsing, suitable for transmission.

Among the Disadvantages: need for a off-line processing.

R.Szeliski [137] [136]

R. Szeliski proposes different ways of computing image warping and he recovers a 3D model depending on the application: 2D planar image mosaicing, partial 3D model recovery and fully 3D model recovery. When recovering a full 3D model the utilized techniques are: volumetric description from silhouette or stereo matching from image sequence.

In volumetric description, the 3D model is recovered from a binary silhouette of an object against its background, local optical flow is computed and converted into sparse 3D point estimates, and the occluding contours of an object are tracked to generate 3D space-curves.

These techniques are suitable to reconstruct an isolated object undergoing known motion. Similar techniques can however be used to solve a more general 3D scene recovery problem where the camera motion is unknown. In particular, it is proposed the projective motion algorithm for determining an object motion based on recovering ”projective depths”.

Among the advantages: possibility for automatically creating large panorama images of arbitrary shape and detail.

Among the Disadvantages: limited 3D rendering.

CohenOr-Rich-Lerner-Shenkar [33]

The paper proposes the use of a textured mapped voxel-based model to represent terrains and 3D objects.

The use 3D voxels-model is proposed because this fits better a high-detailed real-object, such as a real terrain, than a polygonal-model. Terrains are textured from b/w photographs and some objects (e.g. house buildings) are textured by a more detailed texture. The author uses b/w photographs because of the applications on Missile cameras. This leads to generate a less realistic rendering than using colors but it allows for real-time performance.

The system is based on a portable software rendering able to generate photo-realistic images in real-time (on a parallel machine). This performance is due to an innovative rendering algorithm based on discrete optimized ray-casting algorithm, accelerated by ray-coherence and multiresolution transversal.

Among the Advantages: real-time fly-through, portable software rendering, photo-realism.

As mentioned above, some of the research works have mostly focused on one or few aspects involved in realistic visualization of virtual-views. These works are also relevant for model-based rendering. For example, an automatic extraction of textures from a generic video sequence could represents a convenient approach to model-based texture-mapping, as proposed by the research work following summarized.

Ofek-Shilat-Rappoport-Werman [110]

The proposed method focus on automatically deriving realistic 2D textures from video sequences for texture mapping purposes. The term realistic is here used to indicate textures free of disturbing effects such as highlights, reflections, shadows, etc.

The recorded scene is a video-sequence where the object of interest is viewed in different resolutions and different perspectives. A simple 3D model may also be provided by the user to improve system performance. The authors propose a model given by hand from five point at least. The authors also discuss an automatic model generation through a mask.

A multiresolution texture is proposed and for each pixel the color is computed by a weighted average from correspondent pixels in the video sequence. The multiresolution texture is proposed to be exploited for generation of virtual views of recorded scenes.

The proposed approach allows for identification of undesired features like highlights, reflections and shadows. The system is able to recognize the above features by looking at patches which contain very sharp step edges, since these are most likely to occur with depth discontinuities or reflective highlights. The system is then able to conveniently remove the disturbing features from the texture.

The image quality in resulting textures is as high as the original video-stills and so suitable to be used as reference-views. During visualization reference-views can be selected view-dependently based on current observation viewpoint, and the closer reference-view can be used for the mapping.

Among the advantages: suitable for mapping textures on 3D models from video-sequences, suitable for merging texture appearing in different resolutions, efficient storage of the resulting texture in a multiresolution data structure.

The following two methods are related to both model- and image- based rendering, but they represent different implementations. In the first work, the authors propose both the two approaches one at a time, for the same application contexts. In the second more complex approach, the system generates virtual views based on both, a reconstructed geometric model and an image-based texture-mapping (view-dependent).

Blanc-Livatino-Mohr [14] [16]

The authors present a methods allowing for the exploration of a 3D scene based on triangular-mesh model recovered from 2D views. A comparison between the proposed method for virtual view-synthesis based on a sparse match and a view-synthesis based on a dense match, [15], is also proposed, (earlier in [14] and later in [16]).

The 2D views are photographs of a real scene and proposed as reference views. From these references either a projective model ([15]) or a triangular 3D mesh ([14]) is estimated. Hence, virtual views can be generated to allow a user to virtually navigate inside the scene and appreciate the tri-dimensional structure.

The method based on a dense match uses point reprojection. This method starts with a dense matching between the reference views and then each matched couple is reprojected using the trilinear relations to generate a new view from arbitrary viewpoints.

The method based on sparse matches uses model-based rendering (based on textured triangles). First, a sparse match (e.g. corner points) between the reference views is computed. Then, a textured filtered triangular mesh is calculated based on an initial Delaunay triangulation. Eventually, new views are synthesized by a model-based rendering.

The realistic effect in the first method is due to the fact that each pixel is directly re-projected from the real references views to the virtual view. In the second method the realistic effect is due to the fact that triangles are filled in with the texture from the reference views.

Among the advantages: no camera calibration, no 3D-model of the scene with the first method (a projective model is enough), fast synthesis, applications to high-rate video compression (only the references views and the displacement of the camera need to be transmitted and transmitted data does not depend on image size), fast sparse matching and no "holes" in the synthesized views with the second method as well as a mostly automated fast and realistic scene modeling.

Among the disadvantages: not real-time matching phase because of the computed dense-matching (first method), "holes" in the synthesized images arising from unmatched points and from adjacent pixels in the reference views which are not adjacent in the synthesized view (first method), more than two reference views are needed because of the trilinear tensor (first method), false matches are exacerbated if visualization is required from viewpoints distant from the original viewpoint (second method).

Debevec-Taylor-Malik [43]

The system developed by the authors use photographs and an approximate geometry to create and render realistic models of architectures.

The system requires only a small number of photographs, i.e. fews different views of the object, and a few indications to specify an approximate geometry and rough correspondences between the photographs. A method for photogrammetric modeling implemented on an interactive modeling program (called "Façade") is proposed for this purpose.

The model is then refined by means of proposed *model-based stereo*, which exploits estimated geometry (a coarse object model) and epipolar relations, to match stereo views on a wide baseline. In particular, by re-projecting one image of the stereo pair from the other image viewpoint. In this way, the foreshortening problem for the wide baseline stereo pair is eliminated and the stereo reconstruction can be done more robustly.

When an object is rendered, this is textured by proposed view-dependent texture-mapping technique which interpolates textures coming from photographs which are closer to current viewpoint. In particular, the interpolation is performed using a geometric model to determine which pixel from each input image corresponds to the desired ray in the output image. Among the corresponding rays, those that are closer in angle to the desired ray are weighted to make the greatest contribution to the interpolated result.

A method for averaging textures of neighboring regions in case when regions are textured from different image-sources is also proposed, in order to avoids seams and abrupt transitions of textures.

A view-dependent texture mapping is later proposed in [44] to further reduce the computational cost and have smoother blending, by means of visibility processing, polygonal view-maps, and projective texture-mapping.

Among the Advantages: sparse set of reference images, wide-baseline stereo matching, realistic response.

Among the Disadvantages: projective texture mapping involves expensive computation, the visibility problem needs to be addressed, texture seams may arise.

Class of Approaches: explicit-geometry rendering [130], volumetric reconstruction [54].

A.3 Summary

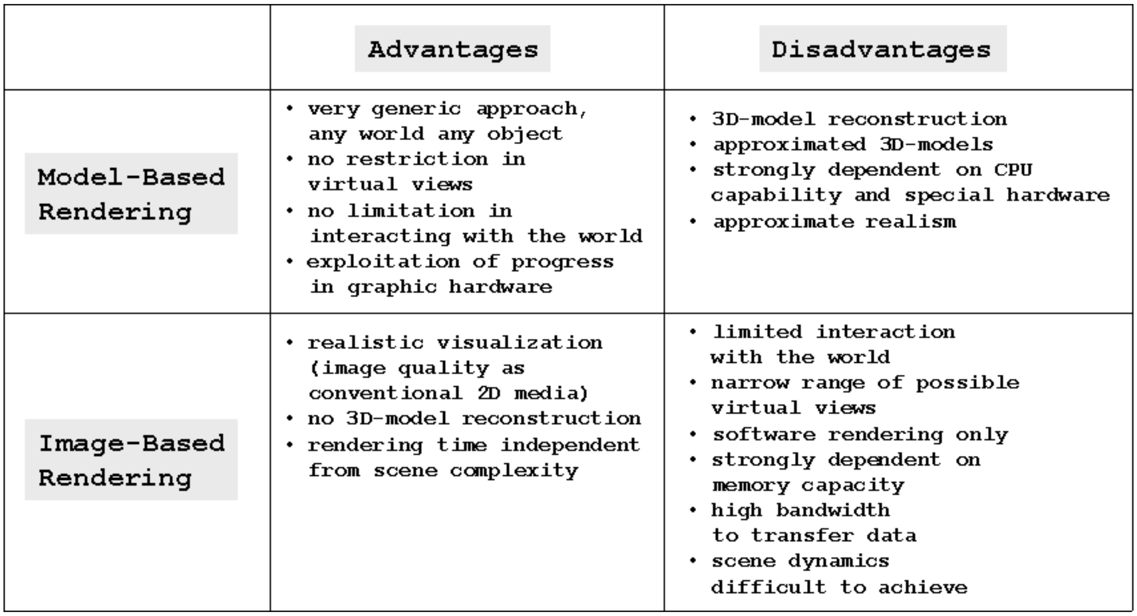
Image-domain approaches emphasize the role of photographs or still-video in order to provide realism to computer generated images. This methodology uses 2D photographic images instead of 3D geometrical models. Realistic virtual views as they are seen from a virtual camera can be generated by an image-based rendering algorithm without any tedious reconstruction of 3D models. Views generated using this method have an advantage compared to those created using the geometrical model-based method, generating the same image quality is much easier. In addition, views generation is relatively easy following "preparation" of the 2D images.

In general image-domain approaches need less computation resource than 3D model-based approaches and the produced image quality is as good as conventional 2D media. However, interaction with the world is limited and they need larger amount of data space, because they have to handle redundant data. Huge amount of data space represents a trade-off for making application involving networks, since a high bandwidth is required to share an image-based virtual world. Also, image-based approaches limit supported virtual views to a "narrow range" and scene is constrained to convex and not occluded objects.

Model-based approaches, on the other hand, are very generic, capable of generating any world and object by using a geometrical model from the beginning. Users encounter no limitations in interacting with the world. Model-based approaches make larger the range of the possible virtual views, and faster the rendering process since they can exploit hardware rendering provided on nowadays graphics-workstations.

Depending on the details and fidelity of the recovered model, model-based methods can yield realistic images. In particular, a realistic synthesis depends on: accuracy of the geometric model of the objects, textures, object surface properties, illumination simulations, rendering algorithms, etc. Pre-acquired information, heuristics and additional effects can also be integrated. A summary of main advantages and disadvantages of the two classes of approaches is presented in figure A.3.

As it comes out from the state of the art, a realistic image synthesis of virtual environments is a large field of applications not yet generally solved. The success of some approaches mainly depends on the application and application constraints. However, from the many different techniques proposed, it is possible to gain a general idea on what approach and on what technique may better fit the individual application dependencies. The parameters playing an important role are:



	Advantages	Disadvantages
Model-Based Rendering	• very generic approach, any world any object • no restriction in virtual views • no limitation in interacting with the world • exploitation of progress in graphic hardware	• 3D-model reconstruction • approximated 3D-models • strongly dependent on CPU capability and special hardware • approximate realism
Image-Based Rendering	• realistic visualization (image quality as conventional 2D media) • no 3D-model reconstruction • rendering time independent from scene complexity	• limited interaction with the world • narrow range of possible virtual views • software rendering only • strongly dependent on memory capacity • high bandwidth to transfer data • scene dynamics difficult to achieve

Figure A.3: The table summarizes advantages and disadvantages of image- and model- based rendering techniques.

* real-time performance;

* static/dynamic environment;

* convexity/concavity of the scene objects;

* object visibility and mutual occlusions;

* range of perspective transformations required;

* level of fidelity required.

Several years ago the image-based approaches were unthinkable to propose since memory devices and high speed data links costed so much. Then amazing advances in semiconductor technologies, including reduction of memory device cost, made possible to explore such methodology.

A remarkable progress has also been reached by computer graphics technology, and by 3D geometric modeling. We currently have many sophisticated 3D graphics tools such as 3D modeling systems, 3D scanning systems, that combined with hours of labor let us generate sophisticated graphics images as seen in movies.

References

[1] E.H. Adelson and J.R. Bergen. *The Plenoptic Function And The Elements Of The Early Vision*, chapter 1. MIT Press, Cambridge, MA, 1991.

[2] C.S. Andersen. *A Framework for Control of a Camera Head.* PhD thesis, Aalborg University, Denmark, 1996.

[3] C.S. Andersen and J.G.M Gonçalves. Determining the pose of a mobile robot using triangulation: a vision based approach. Technical report, EU Joint Research Center, December 1995. No. I.95.179.

[4] K.O. Arras and N. Tomatis. Improving robustness and precision in mobile robot localization by using laser range finding and monocular vision. In *3rd European Workshop on Advanced Mobile Robots (Eurobot)*, pages 177–185, Zurich, Switzerland, Sept. 1999.

[5] P. Aschwanden and W. Guggenbuhl. *Experimental results from a comparative study on correlation-type registration algorithms.* Wichmann, 1992.

[6] I. Asimov. *I, Robot.* Doubleday, 1950.

[7] A. Atiya and G.D. Hager. Real-time vision-based robot localization. *IEEE Transaction on Robotics and Automation*, 9(6):785–800, Dec. 1993.

[8] S. Avidan and A. Shashua. Novel view synthesis in tensor space. In *Conference on computer vision and pattern recognition*, pages 1034–1040, San Juan, Puerto Rico, June 1997.

[9] C. Balkenius. Spatial learning with perceptually grounded representations. *Robotics and Autonomous System, Special Issue on Autonomous Mobile Robots*, 5(3-4):165–175, Nov. 1998.

[10] O. Di Benedetto and S. Livatino. Navigation methodologies for autonomous mobile robots sonar-assisted. Master's thesis, Dept. of Computer Science - University of Pisa, Italy, and Advanced Robotics Technology and System Laboratory (ARTS Lab) - Scuola Superiore S.Anna, Pisa, Italy, April 1993.

[11] R. Benosman and S.B. Kang. *Panoramic Vision: Sensors, Theory, and Applications.* Springer Verlag, 2001.

[12] M. Betke and K. Gurvits. Mobile robot localization using landmarks. *IEEE Transaction on Robotics and Automation*, 13(2):251–263, April 1997.

[13] J. Bjornstrup. Automatic camera calibration using a passive calibration object. Technical report, Aalborg University, Laboratory of Computer Vision and Media Technology, May 1998. http://www.cvmt.dk/˜jorgen/PhD.

[14] J. Blanc, S. Livatino, and R. Mohr. Fast and realistic image synthesis for telemanipulation purposes. In *European Workshop on Hazardous Robotics*, pages 77–83, Barcelona, Spain, November 1996.

[15] J. Blanc and R. Mohr. From image sequence to virtual reality. In E.P. Baltsavias, editor, *ISPRS Intercommission Workshop: From Pixels to Sequences*, pages 144–149, Zurich, Switzerland, March 1995.

[16] J. Blanc and R. Mohr. Towards fast and realistic image synthesis from real views. In *The 10th Scandinavian Conference on Image Analysis (SCIA)*, pages 455–461, Lappeenranta, Finland, June 1997.

[17] T.E. Boult, R.J. Micheals, M. Eckmann, X. Gao, C. Power, and S. Sablac. Omnidirectional video applications. In *8th International Symposium on Intelligent Robotics Systems*, pages 143–151, Reading, United Kingdom, July 2000.

[18] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen. Unstructured lumigraph rendering. In *Computer Graphics (SIGGRAPH'01)*, pages 425–432, 2001.

[19] R. Bunschoten and B. Kröse. 3-d reconstruction from cylindrical panoramic images. In *9th International Symposium on Intelligent Robotics Systems*, Touslouse, France, July 2001.

[20] W. Burgard, A.B. Cremers, D. Fox, D. Hahnel, G. Lakemeyer, D. Shultz, W. Steiner, and S. Thrun. Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, 00:1–53, May 1999.

[21] K. Capek. *R.U.R (Rossuman's Universal Robots)*. 1921.

[22] S. Carlsson. Relative positioning from model indexing. *Image and Vision Computing*, 12(3):179–186, April 1994.

[23] U. Castellani and S. Livatino. Scene reconstruction: Occlusion understanding and recovery. In Robert B. Fisher, editor, *CVonline: The Evolving, Distributed, Non-Proprietary, On-Line Compendium of Computer Vision*. School of Informatics, University of Edinburgh, December 2001. http://www.dai.ed.ac.uk/CVonline.

[24] U. Castellani, S. Livatino, and R.B. Fisher. Improving environment modelling by edge occlusion surface completion. In *1st International Symposium on 3D Data Processing Visualization and Transmission (3DPVT)*, Padova, Italy, June 2002.

[25] N.L. Chang and A. Zakhor. Arbitrary view generation for three-dimensional scenes from uncalibrated video cameras. *Speech and Signal Processing*, 1995.

[26] N.L. Chang and A. Zakhor. View generation for three-dimensional scenes from video sequences. *IEEE Transaction on Image Processing*, 6(4):584–598, April 1997.

[27] R. Chatila and J.P. Laumond. Position referencing and consistent world modeling for mobile robots. In *IEEE International Conference on Robotics and Automation*, pages 138–145, March 1985.

[28] S. Chen and L. Williams. View interpolation for image synthesis. In *Computer Graphics (SIGGRAPH'93)*, pages 279–288, August 1993.

[29] S.E. Chen. Quicktime vr - an image-based approach to virtual environment navigation. In *Computer Graphics (SIGGRAPH'95)*, pages 29–38, August 1995.

[30] H.I. Christensen, N.O. Kirkeby, S. Kristensen, and L. Knudsen. Model-driven vision for in-door navigation. *Robotics and Autonomous System*, 12:199–207, 1994.

[31] J.C. Clarke. Visual beacons for robot localization. In *Computer Vision and Mobile Robotics Workshop (CVMR)*, pages 9–14, Santorini, Greece, Sept. 1998.

[32] D. Cohen-Or. Model-based view extrapolation for interactive vr web-system. In *Computer Graphics International*, Hasselt, Belgium, 23-27 June 1997.

[33] D. Cohen-Or, E. Rich, U. Lerner, and V. Shenkar. Real-time photo-realistic visual fly-through. *IEEE Transactions on Visualization and Computer Graphics*, 2(3), September 1996.

[34] D. Coombs and K. Roberts. Centering behavior using peripheral vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 440–445, New York City, New York, USA, June 1993.

[35] I.J. Cox and Blanche. An experiment in guidance and navigation of an autonomous robot vehicle. *IEEE Transaction on Robotics and Automation*, 7(2):193–204, 1991.

[36] J.L. Crowley. Matemathical foundations of navigation and perception for an autonomous mobile robot. *L. Dorst editor, Reasoning with Uncertainty in Robotics*, 1996.

[37] S.M. Culhane and J.K. Tsostos. An attentional prototype for early vision. In *European Conference on Computer Vision (ECCV)*, 1992.

[38] J. Davis, S.M. Marschner, M. Garr, and M. Levoy. Filling holes in complex surfaces using volumetric diffusion. In *1st International Symposium on 3D Data Processing Visualization and Transmission (3DPVT)*, Padova, Italy, June 2002.

[39] A.J. Davison. *Mobile Robot Navigation Using Active Vision*. PhD thesis, Robotics Research Group, Department of Engineering Science, University of Oxford, October 1999.

[40] A.J. Davison and N. Kita. Sequential localization and map-building in computer vision and robotics. In *European Conference on Computer Vision (ECCV)*, 2000.

[41] A.J. Davison and D.W. Murray. Mobile robot localization using active vision. In *European Conference on Computer Vision (ECCV)*, 1998.

[42] V.C. de Verdiere and J.L. Crowley. Local apperance space for recognition of navigation landmarks. *Robotics and Autonomous Systems*, 1999.

[43] P.E. Debevec, C.J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Computer Graphics (SIGGRAPH'96)*, pages 11–20, August 1996.

[44] P.E. Debevec, Y. Yu, and G. Borshukov. Efficient view-dependent image-based rendering with projective texture mapping. In *9th Eurographics Workshop on Rendering*, pages 105–116, 1998.

[45] R. Deriche. Using canny's criteria to derive a recursively implemented optimal edge detector. *Computer Vision*, 1:167–187, 1987.

[46] G.N. DeSouza and A.C. Kak. Vision for mobile robot navigation: A survey. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(2):237–267, Feb. 2002.

[47] G. Dudek, P. Freedman, and I. Rekletis. Just-in-time sensing: efficiently, combining sonar and laser range data for exploring unknown worlds. In *International Conference on Robotics and Automation*, volume 1, pages 667–672, 1996.

[48] H.F. Durrant-Whyte and J.J. Leonard. Navigation by correlationg geometric sensor data. In *IEEE/RSJ International Workshop on Intelligent Robot and Systems*, Tsukuba, Japan, September 1989.

[49] C.R. Dyer. Volumetric scene reconstruction from multiple views. In L.S. Davis, editor, *Foundations of Image Understanding*, pages 469–489. Kluwer, 2001.

[50] S. Engelson. *Passive Map Learning and Visual Place Recognition.* PhD thesis, Department of Computer Science, Yale University, 1994.

[51] O. Faugeras. Three-dimensional computer vision: a geometric viewpoint. *MIT Press, Cambridge, Massachusetts*, 1993.

[52] L. Feng, J. Borenstein, and H.R. Everett. Where am i? sensors and methods for autonomous mobile positioning. UM-MEAM-94-21 vol.3, Dept. Mechanical Engineering and Applied Machines, University of Michingan, for Oak Ridge National Laboratory, December 1994.

[53] Fisher R.B. (coordinator), Div. Informatics, University of Edinburgh. The camera project (cad modelling of built environments from range analysis), eu tmr-project. http://www.dai.ed.ac.uk/homes/rbf/CAMERA/camera.htm, 1998-2001.

[54] D. Forsyth and J. Ponce. *Computer Vision - A Modern Approach*, chapter 26, Application: Image-Based Rendering, pages 780–808. Alan Api, 2002.

[55] H. Fuchs, G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S. Lee, H. Farid, and T. Kanade. Virtual space teleconferencing using a sea of cameras. In *First International Symposium on Medical Robotics and Computer Assisted Surgery*, pages 161–167, 1994.

[56] J. Gaspar, E. Grossmann, and J. Santos-Victor. Interactive reconstruction from omnidirectional image. In *9th International Symposium on Intelligent Robotics Systems*, Touslouse, France, July 2001.

[57] P. Gaussier, C. Joulain, S. Zrehen, and A. Ravel. Visual navigation in an open environment without map. In *IEEE International Conference on Intelligent Robots and Systems*, pages 545–550, Sept. 1997.

[58] R.C. Gonzalez and P. Wintz. *Digital Image Processing*. Addison-Wesley publishing company, 1987.

[59] S.J. Gortler, R. Grzeszczuk, R. Szeliski, and M.F. Choen. The lumigraph. In *Computer Graphics (SIGGRAPH'96)*, pages 43–54, New Orleans, August 1996.

[60] G.H. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer Academic Publishers, 1995.

[61] Granum E. (coordinator), CVMT lab., Aalborg University. Being there without going (benogo), eu fet-project. http://www.benogo.dk, 2002-2005.

[62] R. Greiner and R. Isukapalli. Learning to select useful landmarks. *IEEE Transaction on Systems, Man, and Cybernetics, Part B: Cybernetics*, 26(3):437–449, June 1996. Special Issue on Learning in Autonomous Robots.

[63] I. Hallmann and B. Siemiatkowska. Artificial landmark navigation system. In *9th International Symposium on Intelligent Robotics Systems*, Touslouse, France, July 2001.

[64] O. Hansen. *On the use of Local Symmetries in Image Analysis and Computer Vision*. PhD thesis, Aalborg University, Denmark, 1992.

[65] R.M. Haralik. Propagating covariance in computer vision. In *12th International Conference on Copmuter Vision and Pattern Recognition*, pages 493–498, Jerusalem, Israel, Oct. 1994.

[66] R. Hartley. In defence of the 8-point algorithm. In *Fifth international conference on computer vision (ICCV'95)*, pages 1064–1070, June 1995.

[67] M. Hashima, F. Hasegawa, S. Kanda, T. Maruyama, and T. Uchiyama. Localization and obstacle detection for a robot for carrying food trays. In *IEEE International Conference on Intelligent Robots and Systems*, pages 345–351, Sept. 1997.

[68] J.B. Hayet, F. Lerasle, and M. Devy. A visual landmark framework for indoor mobile robot navigation. In *IEEE International Conference on Robotics and Automation*, Washington DC, USA, May 2002.

[69] P.S. Heckbert. Fundamentals of texture mapping and image warping. Master's thesis, Dept. of EECS, UCB, June 1989. Technical Report No. UCB/CSD 89/516.

[70] M. Hirose. Image-based virtual world generation. *IEEE Multimedia*, 4(1):27–32, March 1997.

[71] Hogg D. (coordinator), Dept. Computer Studies, University of Leeds. The resolv project (reconstruction using scanned laser and video), eu-project. http://www.scs.leeds.ac.uk/resolv, 1995-1999.

[72] P. Jensfelt. Localization using laser scanning and minimalistic models. Licentiate thesis. Royal Institute of Technology, Stockholm, Apr. 1999.

[73] P. Jensfelt. *Approaches to Mobile Robot Localization in Indoor Environments.* PhD thesis, Royal Institute of Technology, Stockholm, June 2001.

[74] P. Jensfelt and H. Christensen. Laser based position acquisition and tracking in an indoor environment. In *IEEE International Symposium on Robotics and Automation*, volume 1, pages 331–338, Saltillo, Coahuila, Mexico, 1998.

[75] M.R. Kabuka and A.E. Arenas. Position verification of a mobile robot using standard pattern. *IEEE Robotics and Automation*, 3(6):505–516, Dec. 1987.

[76] R.E. Kalman. A new approach to linear filtering and prediction problems. *Transaction ASME J. Basic Engineering*, 82:34–45, 1960.

[77] T. Kanade, P. Narayanan, and P. Rander. Virtualized reality: Concepts and early results. In *IEEE Workshop on Representation of Visual Scenes*, pages 69–76, June 1995.

[78] S.B. Kang. A survey of image based rendering techniques. *VideoMetrics, SPIE*, 3641:2–16, 1999.

[79] S.B. Kang and R. Szeliski. 3-d scene data recovery using omnidirectional multibaseline stereo. In *IEEE Computer society conference on computer vision and pattern recognition*, pages 364–370, June 1996.

[80] R. Klette, K. Schlüns, and A. Koschan. *Computer Vision: Three Dimensional Data from Images*, chapter 2, pages 35–36. Springer-Verlag New York, Incorporated, 1998. http://www.addall.com/Browse/Detail/9813083719.html.

[81] J.M. Leiva, P. Martinez, E.J. Perez, C. Urdiales, and F. Sandoval. 3d reconstruction of a static indoor environment by fusion of sonar and video data. In *9th International Symposium on Intelligent Robotics Systems*, Touslouse, France, July 2001.

[82] J. Leonard and H. Durrant-Whyte. Mobile robot localization by tracking geometric beacons. *IEEE Transaction on Robotics and Automation*, 7(3):376–382, 1991.

[83] J. Leonard and H. Durrant-Whyte. Directed sonar navigation. 1992.

[84] S. Leveau and O. Faugeras. 3-d scene representation as as collection of images and fundamental matrix. Technical Report 2205, INRIA Sophia-Antipolis, February 1994.

[85] M. Levoy and P. Hanrahan. Light field rendering. In *Computer Graphics (SIGGRAPH'96)*, 1996.

[86] Levoy M. (coordinator), Dept. Computer Science, University of Standford. The digital michelangelo project: 3d scanning of large statues. http://graphics.stanford.edu/projects/mich, 2002.

[87] M. Li, H. Schirmacher, and H.P. Seidel. Combining stereo and visual hull for online reconstruction of dynamic scenes. In *IEEE Workshop on Multimedia and Signal Processing*, December 2002.

[88] S. Livatino. Cylindrical epipolar geometry. In Robert B. Fisher, editor, *CVonline: The Evolving, Distributed, Non-Proprietary, On-Line Compendium of Computer Vision*. School of Informatics, University of Edinburgh, July 2003. http://www.dai.ed.ac.uk/CVonline.

[89] S. Livatino. Cylindrical panoramic transfer and appearance prediction for the match of real observations. In Robert B. Fisher, editor, *CVonline: The Evolving, Distributed, Non-Proprietary, On-Line Compendium of Computer Vision*. School of Informatics, University of Edinburgh, July 2003. http://www.dai.ed.ac.uk/CVonline.

[90] S. Livatino. Main approaches in realistic virtual view synthesis. In Robert B. Fisher, editor, *CVonline: The Evolving, Distributed, Non-Proprietary, On-Line Compendium of Computer Vision*. School of Informatics, University of Edinburgh, July 2003. http://www.dai.ed.ac.uk/CVonline.

[91] S. Livatino and D. Hogg. Image synthesis for telepresence. In *European Workshop on Semi-Autonomous Monitoring and Robotics Technology*, Las Palmas, Canary Islands, Spain, 7-10 January 1999.

[92] S. Livatino and C.B. Madsen. Autonomous robot navigation with automatic learning of visual landmarks. In *7th International Symposium on Intelligent Robotics Systems (SIRS)*, pages 419–428, Coimbra, Portugal, July 1999.

[93] S. Livatino and C.B. Madsen. Optimization of robot self-localization accuracy by automatic visual-landmark selection. In *The 11th Scandinavian Conference on Image Analysis (SCIA)*, pages 501–506, Kangerlussuaq, Greenland, June 1999.

[94] S. Livatino and C.B. Madsen. Acquisition and recognition of visual landmarks for autonomous robot navigation. In *8th International Symposium on Intelligent Robotics Systems (SIRS)*, pages 269–279, Reading, United Kingdom, July 2000.

[95] J. Lobo, C. Queiroz, and J. Dias. Vertical world feature detection and mapping using stereo vision and accelerometers. In *9th International Symposium on Intelligent Robotics Systems*, Touslouse, France, July 2001.

[96] C.B. Madsen. A comparative study of the robustness of two pose estimation techniques. *Machine Vision and Applications, special issue on performance characterization*, 9:291–303, Feb. 1997.

[97] C.B. Madsen and C.S. Andersen. Optimal landmark selection for triangulation of robot position. *Robotics and Autonomous Systems*, 23:227–292, 1998.

[98] S. Mann and R.W. Picard. Virtual bellows: Constructing high quality stills from video. In *1st IEEE International Conference on Image Processing*, Nov. 1994.

[99] W.R. Mark, L. McMillan, and G. Bishop. Post-rendering 3d warping. In *Symposium on interactive 3D graphics*, pages 7–16. ACM Press, April 1997.

[100] J. Martin and J.L. Crowley. Comparison of correlation techniques. In *Conference on Intelligent Autonomous Systems (IAS)*, Karlsruhe, Germany, Mar. 1995.

[101] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal region. In *British Machine Vision Conference (BMVC)*, Cardiff, United Kingdom, Sept. 2002.

[102] L. McMillan. Image-based rendering using image-warping - motivation and background. In *Computer Graphics (SIGGRAPH'99), course n.39*, August 1999.

[103] L. McMillan and G. Bishop. Plenoptic modeling: an image-based rendering system. In *Computer Graphics (SIGGRAPH'95)*, pages 39–46, August 1995.

[104] M. Meng and A.C. Kak. Mobile robot navigation using neural network and nonmetrical environment models. *IEEE Control Systems*, pages 30–39, Oct. 1993.

[105] S. Moezzi, A. Katkere, D. Kuramura, and R. Jain. Reality modeling and visualization from mulitple video sequences. *IEEE Computer Graphics and Applications*, 16(6):58–63, November 1996.

[106] S. Moezzi, L. Tai, and P. Gerard. Virtual view generation for 3d digital video. *IEEE Multimedia*, 4(1):18–26, Jan.-Mar. 1997.

[107] H.P. Moravec and A. Elfes. High resolution maps from wide angle sonar. In *IEEE International Conference on Robotics and Automation*, pages 116–121, 1985.

[108] F. Nashashibi, M. Devy, and P. Fillatreau. Indoor scene terrain modeling using multiple range images for autonomous mobile robots. In *IEEE International Conference on Robotics and Automation*, volume 1, pages 40–46, 1992.

[109] J. Neira, M.I. Ribeiro, and J.D. Tardos. Mobile robot localization and map building using monocular vision. In *Symposium on Intelligent Robotics Systems (SIRS)*, 1997.

[110] E. Ofek, E. Shilat, A. Rappoport, and M. Werman. Highlight and reflection-independent multiresolution textures from image sequences. *IEEE Computer Graphics and Applications*, 17(6), March-April 1997.

[111] E. Oliveira and V. Santos. Fibre optics gyroscope evaluation and calibration with mobile robot. In *8th International Symposium on Intelligent Robotics Systems*, pages 281–286, Reading, United Kingdom, July 2000.

[112] Orphanoudakis S.C. (coordinator), Institue of Computer Science, Foundation for Research and Technology - Hellas (FORTH), Crete, Greece. The virgo project (vision-based robot navigation research network), eu tmr-project. http://www.ics.forth.gr/virgo, 1996-2001.

[113] L. Paletta, S. Frintrop, and J. Hertzberg. Robust localization using context in omnidirectional imaging. In *IEEE International Conference on Robotics and Automation (ICRA)*, Seoul, South Korea, May 2001.

[114] S. Peleg and J. Herman. Panoramic mosaics by manifold projection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 338–343, San Juan, Puerto Rico, June 1997.

[115] V.S. Ramachendran. *A.I. and the Eye*, chapter 3, Visual Perception in People and Machines. Wiley and Sons, 1990.

[116] M. Regan and R. Pose. Priority rendering with a virtual reality address recalculation pipeline. In *Computer Graphics (SIGGRAPH'94)*, 1994.

[117] M. Ribo and A. Pinz. A comparison of three uncertainty calculi for building sonar based occupancy grids. In *7th International Symposium on Intelligent Robotics Systems (SIRS)*, pages 235–243, Coimbra, Portugal, July 1999.

[118] B. Rousso, S. Peleg, and I. Finci. Mosaicing with generalized strips. In *DARPA Image Understanding Workshop*, pages 261–264, 1997.

[119] M. Ruggeri, P. Dias, Vitor Sequeira, and J.G.M Gonçalves. Interactive tools for quality enhancement in 3d modelling from reality. In *9th International Symposium on Intelligent Robotics Systems*, Touslouse, France, July 2001.

[120] A. Sabatini, O. Di Benedetto, and S. Livatino. Estimating time-to-crash while tracking an ultrasonic rangefinder. In *International Symposium on Measurement and Control in Robotics (ISMCR)*, Torino, Italy, Sept. 1993.

[121] A.M. Sabatini. Active hearing for external imaging based on an ultrasonic transducer array. In *IEEE/RSJ International Conference on Intelligent Robot and Systems*, pages 829–836, Reyleigh, NC, USA, July 1992.

[122] J. Santos-Victor, G. Sandini, F. Currotto, and S. Garibaldi. Divergent stereo for robot navigation: Learning from bees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 434–439, New York City, New York, USA, 1993.

[123] J. Santos-Victor, R. Vassallo, and H. Schneebeli. Topological maps for visual navigation. In *1st International Conference on Computer Vision Systems (ICVS)*, Las Palmas de Gran Canaria, Spain, Jan. 1999.

[124] C. Schmid and R. Mohr. Local grayvalue invariants for image retreval. *IEEE Transaction on Pattern Analysis and machine Intelligence*, 1999.

[125] S.M. Seitz and C.R. Dyer. Phisically-valid view synthesis by image interpolation. In *Workshop on Representations of Visual Scenes*, 1995.

[126] S.M. Seitz and C.R. Dyer. View morphing. In *Computer Graphics (SIGGRAPH'96)*, pages 21–30, August 1996.

[127] A. Shashua. Algebraic functions for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):779–789, 1995.

[128] A. Shashua and M. Werman. Fundamental tensor: On the geometry of three perspective views. In *IEEE International Conference on Computer Vision (ICCV)*, pages 920–925, 1995.

[129] J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994.

[130] H. Shum and S. Kang. A review of image-based rendering techniques. In *SPIE Int. Conf. on Visual Communication and Image processing*, pages 2–13, 2000.

[131] G. Slabaugh, B. Culbertson, T. Malzbender, and R. Schafer. A survey of volumetric scene reconstruction methods from photographs. In K. Mueller and A. Kaufman, editors, *Volume Graphics 2001, Proc. of Joint IEEE TCVG and Eurographics Workshop*, pages 81–100, Stony Brook, New York, USA, June 2001. Springer Computer Science.

[132] GOLF TAI srl. The wiz. http://www.golf-wiz.com, 2002.

[133] H.W. Stone. Mars pathfinder microrover: A low-cost, low-power spacecraft. In *AIAA Forum on Advanced Developments in Space Robotics*, Madison, Wisconsin, USA, 1996.

[134] M. Störring. Fixation and tracking using active camera with foveated wide-angle lens. Master's thesis, Aalborg University, Institute of Electronic Systems, Laboratory of Image Analysis (LIA), Dec. 1997.

[135] F. Stulp. *Completion of Occluded Surfaces.* PhD thesis, Rijksun Universiteit, Groningen, Holland, 2001.

[136] R. Szeliski. Image mosaicing for tele-reality applications. In *IEEE Workshop on Applications of Computer Vision*, pages 44–53, Los Alamitos, California, 1994. IEEE CS Press.

[137] R. Szeliski. Video mosaics for virtual environments. *IEEE Computer Graphics and Applications*, pages 22–30, March 1996.

[138] D. Tell and S. Carlsson. View based visual servoing using epipolar geometry. In *The 11th Scandinavian Conference on Image Analysis (SCIA)*, pages 63–70, Kangerlussuaq, Greenland, June 7-11 1999.

[139] S. Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artificial Intelligence*, 99(1):21–71, Feb. 1998.

[140] S. Thrun, M. Bennewitz, W. Burgard, A.B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, J. Shulte, and D. Shultz. Minerva: A second generation mobile tour-guided robot. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1999.

[141] P.E. Trahanias, S. Valissaris, and T. Garavelos. Visual landmak extraction and recognition for autonomous robot navigation. In *International Conference on Intelligent Robots and System (IROS)*, 1997.

[142] E. Trucco and A. Verri. *Introductory Techniques for 3-D Computer Vision*, chapter 7, Stereopsis. Prentice Hall, Inc., 1998.

[143] R.Y. Tsai. A versatile camera calibration technique for high accuracy 3d machine vision metrology using off-the-shelf tv camera and lens. *IEEE Robotics and Automation*, 3(4), Aug. 1987.

[144] T. Tsamura. Survey of automated guided vehicle in japanese factory. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1329–1334, Apr. 1986.

[145] B. Tseng and D. Anastassiou. Compatible video coding of stereoscopic sequences using mpeg-2's scalability and interlaced structure. In *International Workshop on HDTV'94*, Torino, Italy, October 1994.

[146] B. Tullsson. Topics in fmcw radar disturbance suppression. *Radar*, pages 1–5, Oct. 1997.

[147] O. Wijk, P. Jensfelt, and H. Christensen. Triangulation based fusion of ultrasonic sensor data. In *IEEE International Conference on Robotics and Automation*, volume 4, pages 3419–3424, Leuven, Belgium, May 1998.

[148] N. Winters and J. Santos-Victor. Omni-directional visual navigation. In *7th International Symposium on Intelligent Robotics Systems (SIRS)*, pages 419–428, Coimbra, Portugal, 1999.

[149] Y. Yeshurun and E.L. Schwartz. Cepstral filtering on a columnar image architecture: A fast algorithm for binocular stereo segmantation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):759–767, July 1989.

[150] D.C.K. Yuen and B.A. MacDonald. Considerations for the mobile robot implementation of panoramic stereo vision system with a single optical centre. In *Image and Vision Computing*, pages 335–339, Auckland, New Zealand, Nov. 2002.

[151] Z. Zhang, R. Deriche, O. Faugeras, and Q.T. Luong. A robust technique for matching two uncalibrated images trough the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78(1-2):87–119, 1995.

[152] P. Zingaretti and A. Carbonaro. Route following based on adaptive visual landmark matching. *Robotics and Autonomous System, Special Issue on Autonomous Mobile Robots*, 25(3-4):177–184, Nov. 1998.

[153] P. Zingaretti and A. Carbonaro. Learning to acquire and select useful landmarks for route following. In *3rd European Workshop on Advanced Mobile Robots (Eurobot)*, pages 160–168, Zurich, Switzerland, Sept. 1999.

[154] P. Zingaretti, A. Carbonaro, and P. Puliti. Image segmentation for appearance-based self-localisation. In *11th International Conference on Image Analysis and Processing (ICIAP)*, Los Alamitos, USA, 2001.

Computer Vision & Media Technology Laboratory

THE COMPUTER VISION AND MEDIA TECHNOLOGY LABORATORY (CVMT) IS PART OF THE INSTITUTE OF HEALTH SCIENCE AND TECHNOLOGY. CVMT WAS FOUNDED IN 1984 AS THE LABORATORY OF IMAGE ANALYSIS (LIA). THE MAIN RESEARCH AREAS OF THE LABORATORY ARE COMPUTER VISION, MULTIMEDIA INTERFACES, VIRTUAL REALITY, AUGMENTED REALITY SYSTEMS, AND AUTONOMOUS SYSTEMS AND AGENTS.

CVMT HAS ESTABLISHED RESEARCH COOPERATION WITH MORE THAN 30 INSTITUTIONS IN 16 DIFFERENT COUNTRIES.

CVMT IS HEADED BY ITS FOUNDER PROFESSOR ERIK GRANUM.

CVMT
Computer Vision
& Media Technology Laboratory
Aalborg University
Niels Jernes Vej 14
DK-9220 Aalborg
Denmark

Telephone: +45 9635 8789
TeleFax: +45 9815 2444

E-Mail: info@cvmt.dk
URL: http://www.cvmt.dk