5-2008

Long Range Automated Persistent Surveillance

Yi Yao
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a dissertation written by Yi Yao entitled "Long Range Automated Persistent Surveillance." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Electrical Engineering.

Mongi A. Abidi, Major Professor

We have read this dissertation and recommend its acceptance:

Besma R. Abidi, Paul B. Crilly, Hairong Qi, Hamparsum Bozdogan

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Yi Yao entitled “Long Range Automated Persistent Surveillance”. I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Electrical Engineering.

Mongi A. Abidi, Major Professor

We have read this thesis
and recommend its acceptance:

Besma R. Abidi

Paul B. Crilly

Hairong Qi

Hamparsum Bozdogan

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the
Graduate School

(Original signatures are on file with official student records.)

Long Range Automated Persistent Surveillance

A Dissertation Presented for the
Doctor of Philosophy Degree
The University of Tennessee, Knoxville

Yi Yao
May 2008

Acknowledgement

First and foremost, I am deeply indebted to my parents, Zhongqing Jin and Qi Yao, who have always encouraged me to pursue higher education. To my husband, Yanzhen Li, who has always supported me during this journey. To my sister, Lan Jin, and her family for their love.

I would like to thank my advisor, Dr. Mongi Abidi, who has made me who I am and who I will be in my professional career. His willingness to support my work and his guidance throughout my studies has allowed me to develop my abilities to think through the problems and skills to find the answers. I thank him for that opportunity. Special thanks indeed go to Dr. Besma Abidi, whose regular technical comments provided me valuable guidance throughout my research. Also, I would like to thank Dr. Paul Crilly, Dr. Hairong Qi, and Dr. Hamparsum Bozdogan. Their advice and counsel have been of equal importance. I greatly appreciate their time and input to this dissertation.

Within the IRIS group, to Dr. David Page and Dr. Andreas Koschan I express my sincerest gratitude for the opportunity and many conversations that have had tremendous impacts on my research. I owe many thanks to my fellow graduate students, Hong Chang, Chunghao Chen, Harishwaran (Hari) Hariharan, and Sreenivas Rangan. The administrative staff, Justin Acuff, Vicki Courtney-Smith, Kim Cate, Diane Strutz, and Doug Warren for their help and support. Thank you all.

Abstract

This dissertation addresses long range automated persistent surveillance with focus on three topics: sensor planning, size preserving tracking, and high magnification imaging.

For an automated and persistent surveillance, a sufficient overlap of the camera's field of view should be reserved so that camera handoff can be executed successfully before the object of interest becomes unidentifiable or untraceable. We design a sensor planning algorithm that not only maximizes coverage but also ensures uniform and sufficient overlapped camera's field of view for an optimal handoff success rate. This algorithm works for environments with multiple dynamic targets using different types of cameras. Significantly improved handoff success rates are illustrated via experiments using floor plans of various scales.

Size preserving tracking automatically adjusts the camera's zoom for a consistent view of the object of interest. Target scale estimation is carried out based on the paraperspective projection model which compensates for the center offset and considers system latency and tracking errors. A computationally efficient foreground segmentation strategy, 3D affine shapes, is proposed. The 3D affine shapes feature direct and real-time implementation and improved flexibility in accommodating the target's 3D motion, including off-plane rotations. The effectiveness of the scale estimation and foreground segmentation algorithms is validated via both offline and real-time tracking of pedestrians at various resolution levels.

Face image quality assessment and enhancement compensate for the performance degradations in face recognition rates caused by high system magnifications and long observation distances. A class of adaptive sharpness measures is proposed to evaluate and predict this degradation. A wavelet based enhancement algorithm with automated frame selection is developed and proves efficient by a considerably elevated face recognition rate for severely blurred long range face images.

Table of contents

List of figures

List of tables

1 Introduction

Safety and security in public locations have received intensive attention in recent years and especially after the tragedy of 9/11. For security purposes, such locations often rely on camera systems for activity monitoring, threat assessment, and situational awareness. With the increased scale and complexity involved in most practical surveillance situations, it is almost impossible for any single camera (either omnidirectional or PTZ) to fulfill the tracking and monitoring tasks with an acceptable degree of continuity and/or reasonable accuracy. As a result, systems with multiple cameras have entered into play and found extensive applications. The question of how to place multiple cameras to accomplish the given tasks arises naturally, followed by the question of how to manage multiple cameras automatically in real time so that the objects of interest can be monitored continuously. In addition, as the required system's intelligence level increases, object recognition and activity understanding are performed for threat assessment and situational awareness. This introduces extra resolution requirements and the second question raised above then becomes: how to manage multiple cameras automatically in real time so that the objects of interest can be monitored continuously and with the required degree of details. The dissertation work described herewith resolves the aforementioned questions and extends the research to long range surveillance and high magnification video processing.

The remainder of this chapter outlines the motivation for this research in section 1.1. Section 1.2 gives a brief review of the state of the art. The pipeline and contributions of this dissertation are presented in section 1.3. Section 1.4 concludes this chapter with the document organization.

1.1 Motivation

Sensor planning for surveillance systems has received increasing attention in recent years. Cameras are placed to achieve a full or specified coverage of the environment. The performance of camera placement depends on the modeling of the environments and cameras, the design of the objective function representing the given tasks, and the effectiveness of the optimization algorithm used. A large number of ineffective camera arrangements exist in current surveillance systems. Figure 1.1 illustrates the camera arrangement on the third floor of the Electrical Engineering building of the University of Tennessee. It is clear that the existing camera arrangement shown in Figure 1.1(a) does

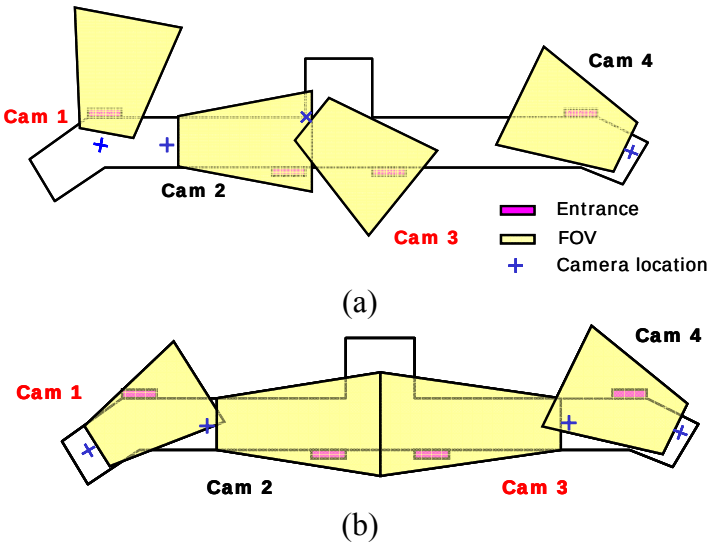
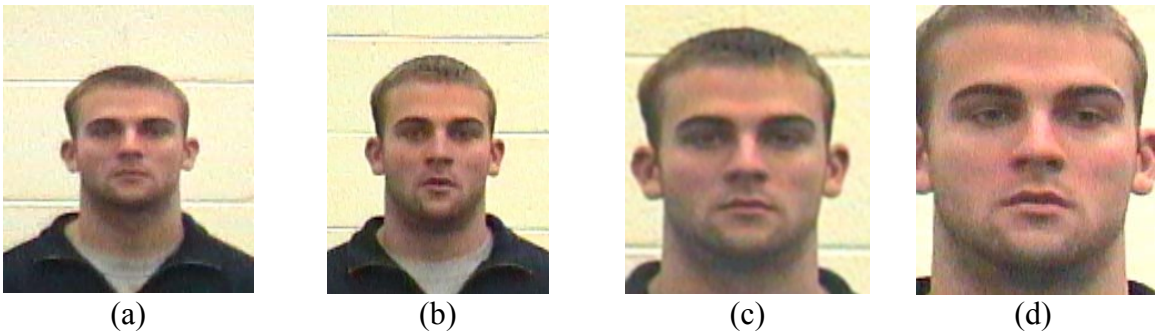


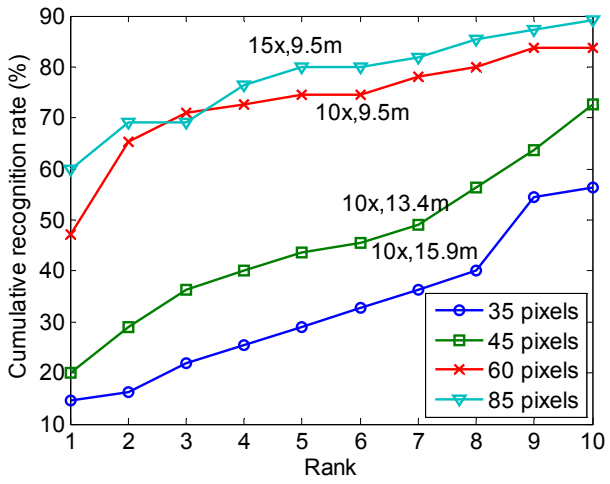
Figure 1.1. Illustration of the need for sensor planning. (a) Sketch of the camera arrangement of an existing surveillance system: the third floor of the Electrical Engineering building, the University of Tennessee, Knoxville. (b) The camera arrangement after sensor planning for coverage maximization. With the same number of cameras of the same models, the coverage is improved substantially without loosing the focus on the entrance areas.

not make full use of the cameras' field of view (FOV), especially for cameras 1 and 3. With the adjusted positions of cameras 1 and 3 shown in Figure 1.1(b), which are obtained via sensor planning for coverage maximization, all entrances are covered with a significantly improved coverage. This verifies the need for a systematic method to optimize the camera placement. Furthermore, the conventional requirements in sensor planning, such as coverage and visibility [Erdem06], alone are unable to ensure a persistent and automated tracking in real-time surveillance. A uniform and sufficient amount of overlap between the FOVs of adjacent cameras should be reserved so that consistent labeling and camera handoff can be executed successfully.

To illustrate the resolution requirement encountered in a surveillance system, we consider face recognition as an example application. Along with illumination and pose, resolution constitutes one of the most decisive factors in face recognition. For a successful recognition, a minimum resolution is required. For instance, a resolution corresponding to an inter-ocular distance of 60 pixels is recommended by FaceIt® [Phillips02]. Figure 1.2 demonstrates the cumulative match characteristics (CMC) curves, obtained by computing the cumulative percentage of correctly recognized probes at various ranks, and manifests the degradation in face recognition rates (FRR) caused by decreased resolution. The CMC measure (CMCM) and rank-one recognition rate, as listed in Table 1.1, are used to evaluate the overall face recognition performance. The CMCM is a quantified measure of a CMC curve and is defined as $Q_{CMC} = \sum_{k=1}^{N_{rank}} C_k / k$, where N_{rank} is the total number of ranks considered and C_k denotes the percentage of



(a) (b) (c) (d)



(e)

Figure 1.2. Face recognition rate vs. face resolution measured by the inter-ocular distance in pixel. (a)-(d) Sample face images with various inter-ocular distances: (a) 35 pixels, (b) 45 pixels, (c) 60 pixels, and (d) 85 pixels. (e) CMC comparison across face resolutions. Gallery images are collected by a Canon A80 camera with a focal length of 114mm and a resolution of 2272×1704 from a distance of 0.5m. Probe images are collected by a Panasonic PTZ camera (WV-CS854) with varying camera zooms (10×~15×) and observation distances (9.5m~15.9m). Database size: 55 subjects.

Table 1.1. Performance comparison across face resolutions based on the CMCM and rank-one recognition rate.

Magnification, distance, resolution	CMCM (%)	Rank-one (%)
10×, 15.9m, 35p	22.3	14.6
10×, 13.4m, 45p	32.1	20.0
10×, 9.5m, 60p	60.2	47.3
15×, 9.5m, 85p	68.2	60.0

probes correctly recognized at rank k [Yao06D]. In our experiments, $N_{rank} = 10$ is used. The overall FRR drops substantially with respect to a reduced face resolution, indicated by a decrease of 45.9% in CMCM as the inter-ocular distance decreases from 85 pixels to 35 pixels.

To achieve and maintain the required resolution on the object of interest, we consider the following two aspects: (1) object tracking algorithms with automatic zoom control to maintain the required resolution while the target is in the camera's FOV and (2) high magnification imaging systems capable of optically achieving the required resolution.

To maintain the required resolution, conventional object tracking, where the camera's pan and tilt angles are adjusted so that the target remains in the camera's FOV, is insufficient. The camera's zoom should be varied automatically so that the target also has a constant or a desired image size regardless of its relative motion and distance with respect to the observing camera. To differentiate it from conventional object tracking, we denote our tracking with automatic zoom control as *size preserving tracking*.

Video tracking systems with automatic zoom control have attracted increasing research interests, due to their added flexibility to interact with changing conditions. The concept and advantages of size preserving tracking are clearly illustrated in Figure 1.3. The image sequence in Figures 1.3(a)-(d) is collected using a constant camera zoom. Beyond a certain distance the target's details are unrecognizable as shown in Figures 1.3(a) and (b) to where a larger zoom is preferable. On the other end as shown in Figure 1.3(d), the target is too close for the camera to properly maintain it in the camera's FOV by panning and/or tilting. Under these circumstances, a smaller zoom is required to enlarge the camera's FOV. As Figures 1.3(e)-(h) depict, with proper zoom control, the target remains in the camera's FOV and its image presents the desired details throughout the sequence.

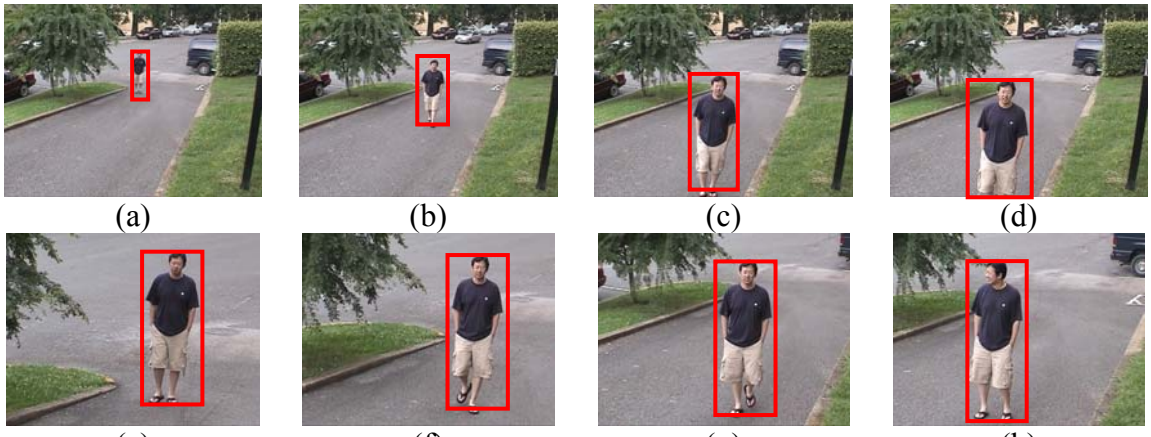


Figure 1.3. Illustration of the need for size preserving tracking based on a pedestrian sequence with (a)-(d) constant camera zoom and (e)-(h) automatically adjusted camera zoom. The red rectangle highlights the tracked target. With proper zoom control, the target remains in the camera's FOV and its image presents the desired details throughout the sequence.

From our experiments, cameras with a 20× zoom capability can monitor an area with a radius of 15m while maintaining the required resolution for a successful face recognition. For a larger surveillance area (radius>50m) or a better resolution, the zoom capability of commercial PTZ cameras is no longer adequate, which necessitates the use of composite imaging systems. Composite imaging systems are traditionally employed in astronomy and wild life monitoring. More recently, the need for such capabilities has extended to near-ground surveillance. To achieve high magnification and long observation distance, we designed a number of composite imaging systems by coupling off-the-shelf scopes (telescopes or spotting scopes) with digital cameras/camcorders and utilized the resulting imaging systems in near-ground and real-time surveillance including object tracking and face recognition. Images with high magnification suffer from various types of degradations, such as increased image noise, severe image blur, and low intensity contrast. In this effort, a comprehensive processing algorithm designed for long range face images is discussed, including frame selection, noise reduction, and facial detail enhancement.

Figures 1.4 and 1.5 illustrate a typical scenario where this dissertation is applicable. Multiple cameras are employed to monitor all the latent activities in the environment. Specified resolution sufficient for identity verification is required when a worker carries valuable assets along the path toward the RFID detector until he or she drops the assets in the storage area. To fulfill all the required monitoring and tracking tasks, sensor planning, size preserving tracking, and high magnification imaging are necessary.

1.2 State of the art

Although multi-camera surveillance systems have resulted in intensive research efforts, most of the existing work remains in solving the problem of consistent labeling, which relates and identifies the projected images of the same target in different cameras. In literature, consistent labeling could be grouped into two categories: feature-based and geometry-based approaches. Feature-based methods search for a match of distinguishing features, such as the color distribution of the tracked objects, and generate correspondences among cameras [Chang01, Kogut01, Nummiaro03, Utsumi04]. In geometry-based algorithms, the trajectory of the tracked object is projected into the world or a reference coordinate system. Consistent labeling then can be established based on the equivalence between objects projected onto the same location [Black01, Cai99, Kelly95, Tan94].

With multiple cameras, surveillance systems need sensor planning. There exist many sensor planning algorithms in literature focusing on such applications as 3D object inspection and reconstruction. Roy *et al*. reviewed existing sensor planning algorithms for 3D object reconstruction [Roy04] and proposed an online scheme using a probabilistic reasoning framework for next-view planning [Roy05]. Yous *et al.* designed an active scheme for the assignment of multiple PTZ cameras so that each camera observes a specific part of a moving object (mainly pedestrian) and achieves the best visibility of the entire object [Yous06]. Wong and Kamel compared the viewpoint

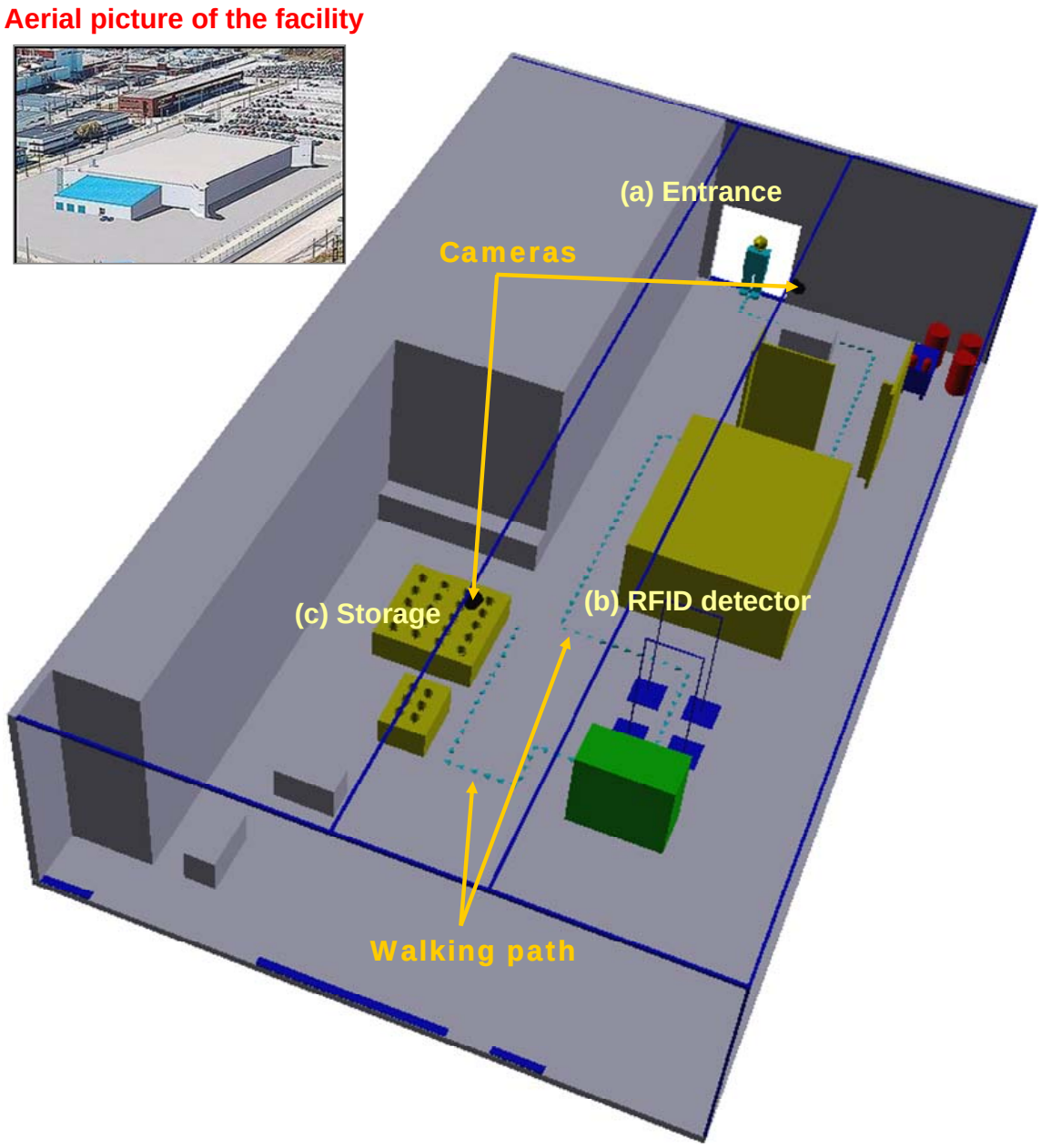


Figure 1.4. Illustration of a typical scenario where this dissertation work is applicable. 3D illustration of the environment to surveil. Sample pictures collected at the specified positions are shown in Figure 1.5. Courtesy of BWXT Y-12.

(a)

(b)

(c)

(d)

(e)

Figure 1.5. Sample pictures collected at the positions specified in Figure 1.4, including (a) entrance, (b) RFID detector, and (c) storage. Detailed view of the (d) RFID detector and (e) glove box and storage areas. Courtesy of BWXT Y-12.

evaluation functions for model based inspectional coverage [Wong04]. Research work was also done on sensor planning for surveillance systems [Cai99, Quereshi05]. Cameras are placed to achieve a full or specified coverage of the environment. A probabilistic camera planning framework with visibility analysis was proposed by Mittal and Davis [Mittal04]. Erdem and Sclaroff defined different types of coverage problems and developed corresponding solutions using PTZ cameras [Erdem06].

Tracking can basically be defined as the search for the optimal matches in consecutive frames. In literature, video tracking schemes only considering pan/tilt control are a well developed research topic. If conventional video tracking schemes are considered as 2D based, the involvement of zoom adds a third dimension. The additional dimension is reflected by either variations in the target's size or additional information about the target's movements in the 3D world coordinates, especially the motion along the camera's optical axis.

The introduction of zoom brings in challenges in three main aspects. (1) The involvement of zoom affects the selection and use of features. For instance, region based methods suffer from the problem of being zoom variant by nature. Contour based methods can compensate for limited degree of deformation but fail when part of the contour falls out of the image, which is commonly encountered during zoom-in operations. (2) A varying zoom imposes extra obstacles and computational burdens on target pursuing. For region based methods, the template must be updated timely or scaled accordingly to keep up with the variations in the target's image size. In point (image corner) based methods, the differentiation between the target movement and background movement is a major concern. (3) The third difficulty has roots in the zoom control itself. The appropriate focal length, capable of compensating for the targets' movement along the camera's optical axis and of producing the desired target image size, has to be determined from a 2D image sequence. In addition, the challenges in practical implementation include the nonlinear and device dependant relation between the system's focal length and zoom control, system delay introduced by mechanical parts and image acquisition, and concerns about system stability.

The most widely used approaches of size preserving tracking are region based methods, where the size, area, and variance of the detected target image are used [Collins03, Hoad95, Kuo02]. Recently, two new trends emerged. One led by Tordoff and Murray [Tordoff00, Tordoff01, Tordoff04] utilizes the concept of structure from motion (SFM) and the other proposed by Fayman *et al*. [Fayman98, Fayman01] is based on the optical flow of the image sequence. Apart from these two algorithms, methods using wavelet transform establish another promising approach [Wei01], where the area of the detected motion blob in the transformed domain is used for zoom control.

Over the last two decades, intensive research work has been conducted in face recognition. Most of the existing work concentrates on scenarios with varying illumination, varying pose, and partial occlusion using still face images or videos collected from a close distance and with a low and constant zoom. Little research attention is paid to face recognition in long range. However, face quality assessment and enhancement algorithms proposed within close range can also serve as references and will be reviewed in the scope of this dissertation.

In face detection and tracking, measurement functions are used to describe the probability of an area being a face image. The term face quality assessment was first explicitly used by Identix [Griffin05], where a face image is evaluated according to the confidence of detectable eyes, frontal face geometry, resolution, illumination, occlusion, contrast, focus, *etc*. Kalka *et al.* applied quality assessment metrics for iris to face images [Kalka06]. Criteria such as lighting (illumination), occlusion, inter-ocular distance (resolution), and image blurriness caused by both out-of-focus and motion are considered. Xiong and Jaynes developed a metric based on bilateral symmetry, color, resolution, and expected aspect ratio (frontal face geometry) to determine whether the current detected face image in a surveillance video is suitable to be added to an on-the-fly database [Xiong03].

Deblurring algorithms are proposed especially for face images by making use of known facial structures. Fan *et al*. incorporated the prior statistical models of the shape and appearance of a face into the formulation of regularized image restoration [Fan03]. A hybrid recognition and restoration architecture was described by Stainvas and Intrator [Stainvas00], where a neural network is trained on both clear and blurred face images. Liao and Lin applied the Tikhonov regularization to Eigen-face subspaces to overcome the algorithm's sensitivity to image noise [Liao05]. Apart from algorithms designed particularly for face images, there exist two major categories of image deblurring techniques, referred to as image sharpening and image restoration by deconvolution. Image sharpening explores image edges or high frequency components to bring out previously invisible details. As for image restoration based on deconvolution, the blurred image is modeled as the original image convolved with a 2D filter. The goal of image restoration is to undo the convolution and in turn eliminate the blur. Unsharp masking using the Laplacian filter is a well-known example of sharpening methods. The classic linear unsharp masking technique suffers from two main drawbacks: sensitivity to noise and overshoot artifacts. Various approaches have been suggested to overcome the aforementioned drawbacks. Many of these schemes are based on the use of nonlinear operators [Ramponi98A, Ramponi98B], where the sharpening operation is controlled by the local activities of the image gradients. Image deconvolution can improve the image's dynamic range and resolve blur simultaneously. Commonly used algorithms are the Lucy-Richardson algorithm, the maximum entropy method, and the Wiener filtering. Regularized deconvolution handles ill-posed problems by adding a regularization term. The Tikhonov [Tikhonov77] and total variation regularization [Chan99] are two such popular choices.

1.3 Contributions

The pipeline of this dissertation work is illustrated in Figure 1.6. An automated and persistent surveillance system using multiple cameras is developed including sensor planning, size preserving tracking, and camera handoff. For long range applications, high magnification imaging systems are employed, which include data acquisition and image

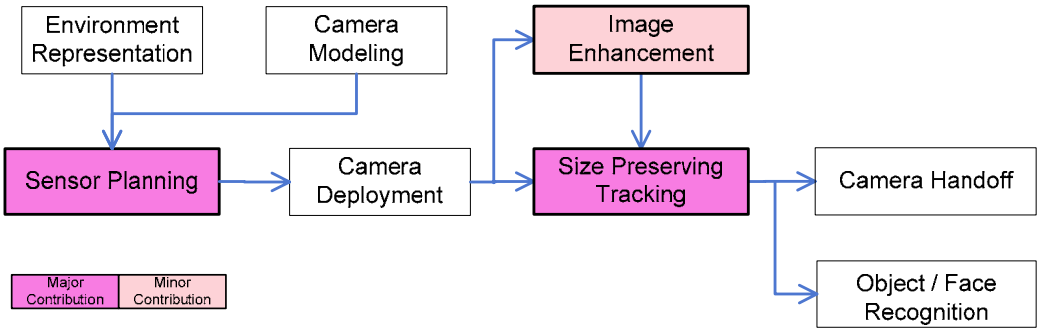


Figure 1.6. The pipeline of our high magnification imaging and video surveillance system.

quality assessment and enhancement. Accordingly, our research contributions are listed as follows.

- **Sensor planning**: most existing camera placement algorithms focus on coverage and/or visibility analysis, which ensures that the object of interest is visible in the camera's FOV. However, visibility alone, a fundamental requirement of object tracking, is inadequate for persistent and automated surveillance. In such applications, a uniform and sufficient overlap between the FOVs of adjacent cameras should be secured so that camera handoff can be executed successfully and automatically before the object of interest becomes untraceable or unidentifiable. From this perspective, our proposed sensor planning method improves existing algorithms by adding handoff rate analysis, which preserves necessary overlapped FOVs for an optimal handoff success rate. In addition, our proposed algorithms also consider multiple dynamic targets where the real-time interaction among moving targets and observing cameras is taken into account via a probabilistic framework.

- **Size preserving tracking:** Tordoff and Murray proposed a scale estimation method based on the weak perspective projection model [Tordoff04]. To account for center offset, the distance between the center of mass of the target's image and the camera's principal point, the paraperspective projection model, a more advanced affine projection model, is utilized and the corresponding scale estimation algorithm is proposed. Furthermore, based on the reconstructed structure, affine shapes of the target are derived and projected into the image to separate the foreground from the background. The resulting segmentation features a fast implementation with linear computations, is robust to off-plane rotation, and allows for deformation to accommodate newly emerged views automatically.

- **Quality assessment and enhancement of high magnification images**: adaptive sharpness measures are designed for the evaluation of image quality under high

magnifications. In addition to illumination and pose, magnification blur is identified as an additional major degradation source in long range face recognition. Wavelet based enhancement algorithms are developed to improve facial features and suppress noise simultaneously. The effectiveness of the proposed assessment and enhancement scheme is validated via a significantly improved FRR in comparison with existing backbone enhancement techniques.

1.4 Document organization

The remainder of this dissertation is organized as follows:

- Chapter 2 reviews existing research work relevant to this dissertation, including multi-camera surveillance, size preserving tracking, and sharpness measures.
- Chapter 3 describes our sensor planning algorithm for multi-camera surveillance.
- Chapter 4 discusses our size preserving tracking algorithm with focus on scale estimation and foreground segmentation.
- Chapter 5 presents the camera handoff algorithm for persistent object tracking.
- Chapter 6 covers out quality assessment and enhancement algorithm for high magnification face images.
- Chapter 7 demonstrates our high magnification imaging system with auto-focusing capabilities.
- Chapter 8 concludes this dissertation with a summary of accomplished and future work.

2 Related work

This chapter discusses existing research work in three relevant areas. Section 2.1 discusses multi-camera surveillance systems. Size preserving tracking algorithms are addressed in section 2.2. A review of sharpness measures, based on which the proposed quality measure for high magnification face images are designed, is given in section 2.3.

2.1 Multi-camera surveillance systems

According to the environments to be monitored and the tasks to be fulfilled, various types of cameras and their combinations are used in multi-camera surveillance systems. In this section, multi-camera surveillance systems are reviewed according to their camera configuration such as perspective cameras, omnidirectional cameras, binocular cameras, and master/slave dual cameras.

2.1.1 Systems using perspective cameras

The collaboration among multiple cameras includes object matching, data fusion, and camera handoff/switching. According to the object matching strategies used, there exist two popular groups of methods: geometry based and feature based methods. Some of the geometry based methods establish the correspondences according to geometric features transformed to the same space, either the ground plane or a common reference plane. Others make direct use of the 3D information obtained from multiple cameras.

The geometric relation between the cameras and the ground plane is calibrated using a set of known landmark features in [Marcenaro01], based on which the targets' 3D trajectories from different cameras (restricted to the ground plane) are obtained and fused. Target detection and tracking are carried out via background subtraction and color histogram matching. If the target is lost from one camera because of occlusion, its position can be restored using data from other cameras. Black and Ellis generated the targets' 2D traces via background subtraction [Black01]. The correspondences of the 2D traces among different viewpoints (cameras) are matched via a least mean squares minimization process. Based on these correspondences, the targets' 3D trajectories are recovered by intersecting the 3D lines connecting the camera's optical center and the 2D trace points. Kalman filter is then used to smooth the raw 3D trajectory. Lee *et al.*

aligned the scene's ground plane across multiple cameras via matching and fitting of tracked objects to a planar model [Lee00]. Afterwards, the planar alignment matrix is decomposed to recover the 3D relative positions of the camera and the ground plane. Cai and Aggarwal used features such as location, intensity, and geometry to match between images collected by different cameras via a Bayesian probability framework [Cai99]. The correspondences between two adjacent cameras are obtained from epipolar constrains. Homography matrices between different viewpoints are also employed to fuse multiple cameras [Morellas03]. The computation is based on a set of known landmark points.

The target's 3D information is used as a fusion reference by Dockstader and Tekalp [Dockstader01] who proposed an integrating algorithm based on Bayesian network. Their algorithm takes in the 2D observations from multiple cameras and outputs 3D state estimates. The 3D state estimates are then fed into a Kalman filter, producing the targets' final 3D trajectories. These 3D position estimates are used to predict the 2D state for each camera, forming a closed loop system. Cupillard *et al*. [Cupillard02, Cupillard04] made use of 3D features, such as position, width, and height, to match targets in various cameras. The target depth is recovered from multiple views, based on which the targets' 3D models are developed and used for object matching and data fusion.

Although it is affected by changes in illumination and variations in sensor responses, color is still commonly used in feature based methods. Compared with geometry based methods, color based methods suffer from low accuracy. Generally speaking, the color histogram of the detected target is used to search for the optimal matches in different cameras [Kogut01, Nummiaro03]. To overcome the aforementioned weakness, several modifications are proposed, such as adaptive color histogram and multiple color histograms.

Hybrid methods are also exploited in literature. Chang and Gong [Chang01] developed geometry based modalities, including epipolar constrains, homograhpy, landmark points, and feature based modalities, such as the apparent height and apparent color. A Bayesian network takes in these features and infers the correspondence of objects across cameras. In [Utsumi04], a 3D color model is established. The projected image of this 3D color model is compared with the detected target so that the target is labeled consistently across different cameras. A survey regarding visual surveillance of object motion and behaviors can be found in [Hu04].

2.1.2 Systems using master/slave dual cameras

The combination of omnidirectional and PTZ cameras, referred to as the dual camera system, is another popular choice for multi-camera surveillance systems. In a dual camera system, the omnidirectional camera detects the target's motion and provides the PTZ camera with the target's geo-location. The PTZ camera is then directed to the target and keeps tracking it. Meanwhile, the omnidirectional camera keeps monitoring new latent activities and fulfils supplementary object tracking when necessary.

Cui *et al*. used background differencing and radial profile for target detection and tracking [Cui98]. The geometric correspondences between the omnidirectional and PTZ

cameras are fitted into a polynomial with a degree of three. Confidence coefficients are assigned to tracking decisions from both cameras. The final tracking follows the one with the higher confidence coefficient. In so doing, tracking ambiguity and occlusion can be resolved and hence an improved tracking accuracy is achieved.

Scotti *et al*. paid more attention to the discussion on the omnidirectional camera's nonuniform resolution and its geometric relation with the PTZ camera [Scotti05]. The target's color, shape, and position are selected as tracking features. The omnidirectional camera performs as a secondary tracker and becomes active only when the PTZ camera loses its target.

In [Lin03], targets are detected via background subtraction and traced via a Kalman filter. The geometric correspondences are obtained by first mapping the image coordinates of the omnidirectional camera to a reference coordinates corresponding to the PTZ camera's zero position (zero pan and tilt angles). Afterwards, the geometric correspondences are transformed to the PTZ camera's current position via rotation.

2.1.3 Systems using other types of cameras

Binocular omnidirectional cameras are used in [Peixoto98, Peixoto00, Yagi02]. The system proposed in [Peixoto98, Peixoto00] detects targets based on background differencing. The target's motion is restricted in the ground plane and hence can be computed from omnidirectional images without ambiguity. Object tracking is performed using Kalman filter. Yagi and Yachida used optical flow to initialize the foreground region [Yagi02]. The histograms of the background and foreground regions are obtained and the radial profile is computed for object tracking.

Morita *et al*. utilized multiple omnidirectional cameras [Morita03]. With the use of multiple omnidirectional cameras, the target's position can be determined more precisely. In the serial work of Zhu *et al*. [Zhu00], two omnidirectional cameras are used to track and recover the target's 3D motion via panoramic stereo. For both cameras, the target's motion is detected by background subtraction and quasi-connected region grouping.

2.2 Size preserving tracking

Our classification of various size preserving tracking algorithms is based on the type of features used and the underlying mathematical framework. The reviewed algorithms are divided into five categories: (1) region based, (2) image corner based, (3) wavelet based, (4) hybrid and other methods including the image velocity based approach, and (5) target depth based methods.

2.2.1 Region based methods

Region based algorithms are inherently zoom variant. To account for changes in the target's image size, additional parameters must be introduced. However, region based

algorithms only consider the 2D image plane and disregard the target's motion in the 3D world coordinates. Thus, even if additional parameters accounting for the change in the target's image size are obtained, they will be restricted to the 2D image plane and usually not able to produce accurate zoom controls. Despite their drawbacks, region based approaches are still competitive due to their relatively low computational complexity. Different quantities are extracted from the region of interest (ROI), such as the area or size [Hager03, Hoad95, Kim03, Kuo02] and the variance [Mirmehdi97].

In the work of Hoad and Illingworth, the camera's zoom is adjusted so that the bounding box of the detected target occupies 90% of the whole image [Hoad95]. A *zoom factor* is defined as the ratio between the current and the required target image sizes. A lookup table devised from lens calibration is used to convert the *zoom factor* into a zoom motor control. A real-time algorithm for tracking human heads is discussed in [Kuo02]. The proposed algorithm is based on an elliptical head tracker [Bircheld97], which generates an ellipse with varying sizes and locations tracing the head's movements. The camera's focal length is adjusted according to the ratio of the desired and current ellipse sizes. Kim *et al.* [Kim03] made use of the area of ROI, A_{ROI}, obtained from color segmentation for automatic zoom control. They defined two experimentally selected limiting values, denoted as *Tele* and *Wide*. If A_{ROI} is greater than *Wide*, the camera's lens turns wide for zooming out and the camera's lens zooms in if A_{ROI} is smaller than *Tele*.

The mechanism of zoom control based on the detected image size is straightforward. However, the resulting zoom control is not precise since the simple linear relation between the target's image size and the camera's focal length is a high abstraction and simplification of the actual projective imaging process. The deficiency of this type of methods is inherent. Thus they are only applicable to cases where accuracy is not crucial. Moreover, the algorithms discussed in [Hoad95] and [Kim03] are device dependent. The lookup table is obtained from pre-calibration, and the parameters *Wide* and *Tele* are derived from experiments. When different cameras, even of the same make, are used, the system should be re-calibrated.

Mirmehdi *et al.* [Mirmehdi97] proposed a zoom initialization scheme, where the goal is to zoom in onto the target so that the target fills up almost the whole image. By assuming a homogeneous background, it is shown that the target's image size is maximized when the image variance is maximized. From this observation a closed loop control system is developed, where the image variance is monitored. The zoom-in operation stops when the image variance starts decreasing. The proposed scheme is efficient in initializing the system's focal length. Nevertheless, it is only able to carry out zoom-in operations by maximizing the image variance. Moreover, the resulting target's image size may not be visually suitable or even may be over-zoomed.

Lindeberg initiated research work using a Gaussian kernel and its derivatives as a basic tool for analyzing structures at different scales [Lindeberg94A, Lindeberg94B]. Based on his fundamental framework, features are detected through a staged filtering parameterized by Gaussian kernels [Lowe99]. These features define stable points in the *scale-space*. *Image keys* are created, which allow for local geometric deformations by representing blurred image gradients in multiple orientation planes and at multiple scales. These *image keys* are in turn used for object tracking and recognition by searching a least

squares solution for the unknown model parameters. The proposed scheme is invariant to translation, scaling, and rotation, and partially invariant to intensity changes and affine projection.

Collins adapted Lindeberg's scale selection methods to the problem of selecting kernel scale for mean-shift blob tracking [Collins03]. Two interleaved mean-shift procedures are employed to search for the optimal position and scale. The blob features at various scales can be detected as points in the *scale-space* that are local maxima both spatially and in scale.

2.2.2 Image corner based methods

Tordoff and Murray showed that during tracking, the camera's zoom acts as a gain between the scene dynamics and tracking errors, providing a trade-off between maximizing resolution and minimizing tracking error [Tordoff03]. Intuitively, when the target is moving fast, more camera's FOV is needed to keep up with its movement. On the other hand, when the target's speed is slow, we have the freedom to zoom in while maintaining satisfactory tracking. Therefore, the maximum allowable focal length can be determined based on the tracking error. Using a zoom invariant Kalman filter, the camera's zoom can be controlled based on the tracking error, in particular the variance of the tracking error, using two criteria. First, the tracking error must remain within a threshold, and second, the resolution should be maximized. One major concern of this method lies in that the system performance depends on the accurate estimation of the system delays. Tordoff and Murray deliberated on the derivation of system delays and their impact on the overall system performance.

To retrieve scale information, Wei and Badawy estimated the inter-image affine transformation from point correspondences [Wei03]. Compared with Tordoff and Murray's approaches [Tordoff04], this algorithm is simplified in two aspects. (1) Corners are detected within a bounding box, which avoids foreground and background segmentation. (2) Image correspondences are confined to inter-image affine transformation. The second simplification suggests that this approach ignores the target's movements in the 3D world coordinates and is actually a 2D image based method. Three tracked points from the moving target are used as the basis for the inter-image affine transformation. The location of the bounding box is obtained following this affine transformation and the target scale is estimated based on the size of the bounding box.

Shah and Morrell incorporated zoom control into tracking algorithms based on particle filters [Shah04]. The camera's zoom is adjusted so that a given percentage (90% is suggested by Shah and Morrell) of the projected particles fall onto the image plane. Compared with the algorithm described by Tordoff and Murray [Tordoff03], the assumptions of Gaussian distribution and linear transition required by Kalman filter are not necessary. Thus the particle filter based methods are capable of representing more complicated motions.

Hatano and Hashimoto introduced a potential function to evaluate the changes in the detected image corners, especially the corners on the target's edges [Hatano03]. This potential function actually measures the differences in both position and size between the

current frame and a reference frame. The goal of tracking and controlling the camera's focal length is to minimize this potential function. The merit of the proposed algorithm is that the camera's focal length is adjusted according to the target's current position relative to the image center. When the target's image is close to the image center, more space is allowed for zooming in. On the other hand, when the target's image is close to the image boundaries, the camera's focal length is decreased to ensure that the target remains in the camera's FOV in spite of the target's current distance to the camera. This avoids the problem of over-zooming. The proposed algorithm relies on targets with regular shapes that can be parameterized. Although a successful application is illustrated to track a ball-shaped target, the extension to targets with arbitrary and non-rigid shapes remains questionable.

2.2.3 Wavelet based methods

The algorithms studied in the previous sections are established in the spatial domain, where operations are performed directly on the pixels in the image. In this section, images are transformed using wavelets first. Algorithms based on foveate wavelet transform and wavelet subspace are reviewed. The transform based algorithms usually yield less computational complexity. Moreover, due to the lack of direct relation with the pixel intensity, a large amount of deformation is allowable.

The central idea of foveate wavelet transform (FWT), a variable resolution technique, is to represent the fovea with higher resolution and the periphery with a lower resolution in a pyramidal representation [Wei01]. Compared with other variable resolution techniques, FWT is shown to have various merits such as linearity preservation, orientation selectivity, and high flexibility. FWT based automatic zoom control is one successful application of FWT.

The implementation of automatic zoom control is primarily based on the motion of the surrounding objects in the FWT representation, which can be approximated by the *foveate potential moving area* (FPMA). The area of FPMA determines both the gaze point and the zoom values. The gaze point is obtained such that the area of FPMA is beyond a predefined threshold, meaning that enough motion exists in the window. Once we are confident that the FPMA captures the motion in the image, the size of the corresponding FPMA is used as an indicator of the size of the moving target. Similar to the approach used in [Kim03], but in the wavelet domain, when the area is small, the camera zooms in. The camera zooms out when the area is large and maintains the current zoom value otherwise.

In the serial work of Krueger [Krueger99, Krueger00], a tracking scheme based on the wavelet subspace is presented. The wavelet subspace is a vector space spanned by a set of wavelets. It is dual to the image subspace. In order to establish tracking, the basic idea is to deform the image subspace so that it imitates the affine deformation of the input image. When tracking is successful, the weight vector in the wavelet subspace should be constant. Compared with tracking in the image space, which usually involves large computations introduced by pixel-wise operations, tracking in the wavelet subspace requires lower dimension and fewer computations. The proposed scheme is robust to

scale, rotation, and translation. Moreover, it is able to handle the deformations caused by facial expressions.

2.2.4 Hybrid and other methods

In practice, some schemes make use of both regions and points to achieve a better performance. Ferrari *et al.* proposed an affine region tracker [Ferrari01], where an affine inter-image transformation obtained from point correspondences is used for zoom control and search window prediction, while region matching is applied for tracking. *Anchor points* along the region's bounding box are retrieved based on which an inter-image affine transform is derived in a similar way as in [Wei03]. The predicted search region is computed from its ancestor using this affine transform. Afterwards, the target region is searched in the neighborhood of the predicted region. Different from pure region based schemes, Ferrari *et al.* employed the affine transform between anchor points to provide the scale information. Point based algorithms can better represent the changes in the target's image size, while region based algorithms are well-developed under the circumstances with a constant focal length. Simple and efficient algorithms are available. The combination of points and regions exploits the competency of both algorithms while avoiding their weaknesses. Scale estimation and object tracking are addressed by point based and region based approaches, respectively.

Schemes based on optical flow are proposed in [Fayman98, Fayman01]. These schemes only consider the target's motion along the camera's optical axis. Under this assumption, changes in the target's image size are properly captured by the image radial velocity. For a constant image size, this image radial velocity should be zero, which establishes the theoretical foundation for estimating the camera's focal length. The proposed method has four major limitations. (1) The target's motion is restricted to be along the camera's optical axis. The computation of the image radial velocity becomes difficult when more complicated motions are involved. (2) The instability of optical flow computation resulting form image noise further deteriorates the system performance. (3) In [Fayman01], it is shown that the proposed algorithm is able to yield exact results only for target points lying on a reference plane. The error introduced by points not lying on the plane, which experience perspective distortions, is another major concern. (4) The proposed algorithm is only able to negate the radial optical flow, which means that it can only maintain the target's image size but not assign one.

In practical situations, where the camera's position and the geometry of the surrounding environments are known and unchanged, proper zoom control can be derived from the camera's tilt angle required by pursuing the target [Kang04]. A lookup table or an approximation function can be constructed with the camera's tilt angle and the desired target image size as the inputs and the zoom control as the output. This saves on computational complexity and processing time. However, this approach highly depends on the actual geometry of the environment and the available tracking precision.

2.2.5 Target depth based methods

In this category, we concentrate on the research work conducted by Tordoff and Murray, which can be further divided into two major types, perspective camera model based and affine camera model based [Tordoff01, Tordoff04]. Points and lines are commonly used features in this category [Reid96, Hayman03]. In comparison, lines invariant to zoom can be located more accurately and show more temporal stability than corner features.

The approaches considered in this category make use of the observation that changes in the target's image size are related to the target's movement along the camera's optical axis. The ratio between the area in the image (*da*) and the area in the scene (*dA*) projected along the ray direction is preserved:

$$\rho = \sqrt{\frac{da\cos\phi}{dA\cos\Phi}} = \frac{f}{\overline{Z}}, \tag{2.1}$$

where ϕ (Φ) represents the angle between the normal direction of the image (scene) and the camera's optical axis, f denotes the camera's focal length, and $\overline{Z}$ is the mean distance between the target and the camera, as shown in Figure 2.1. As a result, a constant target image size is achieved.

Based on the perspective projection model, the target's motion along the camera's optical axis is derived from the movement of the gaze point along the camera's optical axis. The gaze point is defined as the point in the image plane or in the world coordinates at which the camera is supposed to aim. This method is an improvement over the region based methods. The proposed scheme precisely derives the target's shape and movements in the world coordinates, which produces a more accurate estimation. However, the computational complexity is considerably high, involving exhaustive foreground/background segmentation and camera self-calibration. Another restriction lies in that the proposed scheme assumes a planar structure. Although they are common in man-made scenes, planar structures are rare in natural scenes. Moreover, the algorithm is sensitive to image noise.

When the target's relief is comparatively smaller than its distance to the camera, typically smaller than one fifth of its distance to the camera which occurs frequently in tracking and surveillance applications, the affine projection model is sufficient to

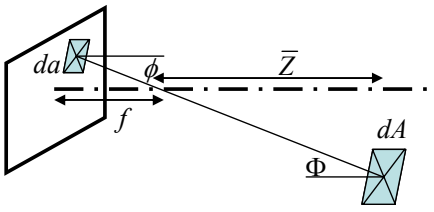


Figure 2.1. Schematic illustration of target depth based methods for size preserving tracking. Courtesy of [Tordoff04].

estimate the changes in the target's image size. In comparison with algorithms based on the perspective projection model, the adoption of the affine projection model substantially reduces computational complexity and improves system stability.

2.2.6 Algorithm comparison

Table 2.1 summarizes and compares the reviewed algorithms. In general, region based methods are not only straightforward in theory but also easy to implement. The existence of well developed region based tracking algorithms constitutes another attractive aspect of these methods. Hence, they are the most popular approaches employed in systems with automatic zoom control. Nevertheless, region based approaches, 2D image based by nature, are of low accuracy in estimating the changes in target scale. The algorithms proposed in the wavelet domain also belong to 2D image based methods and suffer from low accuracy as well. However, they have the advantages of low computations and high tolerance for deformations. Image velocity based methods are straightforward in theory. A direct relation is obtained between the camera's focal length and the image radial velocity. Nevertheless, the difficulties in computing the image velocity accurately, especially when arbitrary target motion is involved, impede their practical application.

In our opinion, target depth based methods are the most promising candidate, primarily due to their superior accuracy. In addition, the simplification from the perspective to the affine projection model, which perfectly adapts to wide area surveillance scenarios, considerably reduces the computational complexity and eliminates the need for planar objects in the scene. The algorithm based on the combined use of regions and image corners is another promising approach, where the advantages of both schemes are explored and where their deficiencies are avoided.

2.3 Sharpness measures

Image artifact, including blur, blocking, and ringing, is commonly encountered in digital image processing. Image quality metrics evaluate the influence of these artifacts. We are interested in image quality measures for our high magnification images. Literature mentions three approaches of image quality assessment: (1) full-reference where the distorted image is compared with its original undistorted image; (2) reduced-reference where the distorted image is compared with a few statistics from its original undistorted image; (3) no-reference where no *a priori* knowledge of the original undistorted image is required. In high magnification imaging systems and in most real-time applications, the original undistorted image is usually not available. Furthermore, the major degradations in high magnification imaging systems are image blur and low contrast. The appropriate candidates of quality assessment for our applications are then

Table 2.1. Comparison of size preserving tracking algorithms.

Algorithms	Advantages	Limitations
Region based methods		
Detected target image size	Straightforward implementation	Linear relation between the target's image size and the camera's focal length
Detected target image variance	Automatically determine the maximum achievable scale	(1) Require homogeneous background (2) Capable of zoom-in operation only
Scale-space	Invariant to rotation and translation	Search in both spatial and scale domains
Image corner based methods		
Tracking error	Low computational complexity	(1) Depend on a zoom invariant Kalman filter (2) Only an upper bound of the zoom range is given
Particle filter	Applicable to general motion and arbitrary noise distributions	High computational complexity
Potential function	Zoom is adjusted by the desired size and the target's image position	Restricted to targets with regular shapes
Wavelet based methods		
Foveate wavelet transform	Automatic latent motion detection	The selection of multiple thresholds
Wavelet subspace	(1)Low computational complexity (2)Ability to handle deformations	Low estimation accuracy
Hilbert transformation	Ability to handle rotation and deformation	A reference template is required
Hybrid and other methods		
Image radial velocity	Direct formula between the camera's focal length and the image radial velocity	(1)Low accuracy in computing optical flow (2)Unable to handle off-plane points
From tilt angle	Essentially a tracking problem	Known surrounding geometry is assumed
Dual camera system	Collaboration between two types of cameras	(1) Known relative geometry is assumed (2) Planar motion is required
Target depth based methods		
Perspective camera model	Accurate	(1) Restricted to planar structure (2) High computational complexity
Affine camera model	(1)Accurate (2)Computationally efficient	The target's relief must be small enough compared with the distance between the target and the camera

narrowed down to sharpness measures, which can be divided into five categories: gradient based, statistics based, autocorrelation based, transform based, and edge based.

2.3.1 Gradient based measures

Grey level differences among neighboring pixels provide a reasonable representation of an image's sharpness. Image gradients obtained by differencing or using high pass filters are abundant in literature. Different forms of gradients can be used [Santos97]: (1) the absolute gradient defined as:

$$S = \sum_{x=1}^{N_{row}} \sum_{y=1}^{N_{col}} | f(x,y+n) - f(x,y) | + | f(x+n,y) - f(x,y) | , \tag{2.2}$$

(2) the squared gradient given by:

$$S = \sum_{x=1}^{N_{row}} \sum_{y=1}^{N_{col}} \sqrt{| f(x,y+n) - f(x,y) |^2 + | f(x+n,y) - f(x,y) |^2} , \tag{2.3}$$

and (3) the maximum gradient formulated as:

$$S = \sum_{x=1}^{N_{row}} \sum_{y=1}^{N_{col}} \max\{| f(x,y+n) - f(x,y) |, | f(x+n,y) - f(x,y) |\} , \tag{2.4}$$

where $f(x,y)$ represents the image intensity, N_{row} (N_{col}) denotes the total number of image rows (columns), and n is the differencing step. The absolute gradient with n=1 is also called the Sum-Modulus-Difference (SMD) and the case with n=2 is commonly referred to as the Brenner measure [Santos97].

The most well-known measure based on high pass filters is the Tenengrad measure [Kroktov89]. The Tenengrad measure is given by:

$$S = \sum_{x=1}^{N_{row}} \sum_{y=1}^{N_{col}} \left[f_x^2(x,y) + f_y^2(x,y) \right], \text{while } \sqrt{f_x^2(x,y) + f_y^2(x,y)} \geq T , \tag{2.5}$$

with the horizontal and vertical gradients, $f_x(x,y)$ and $f_y(x,y)$, obtained using the Sobel filters and T is a threshold. The Laplacian filter is another popular choice [Kroktov89]. The sharpness is defined by:

$$S = \sum_{x=1}^{N_{row}} \sum_{y=1}^{N_{col}} | f(x,y) * h_{Lap}(x,y) | , \text{ while } | f(x,y) * h_{Lap}(x,y) | \geq T , \tag{2.6}$$

where $h_{Lap}(x, y)$ is a Laplacian filter. Choi *et al*. utilized a linear combination of multiple median filters, referred to as the frequency selective weighted median (FSWM) filter [Choi99].

2.3.2 Statistics based measures

Sharp images usually involve large dynamic ranges and scattered grey levels, suggesting a large variance. Two widely recognized sharpness measures are the grey level amplitude and variance. The grey level amplitude, also referred to as the absolute central moment (ACM), is defined as:

$$S = \frac{1}{N_{row}N_{col}} \sum_{x=1}^{N_{row}} \sum_{y=1}^{N_{col}} | f(x,y) - \bar{f} |, \tag{2.7}$$

where $\bar{f} = \frac{1}{N_{row}N_{col}} \sum_{x=1}^{N_{row}} \sum_{y=1}^{N_{col}} f(x,y)$ is the mean grey level. The grey level variance follows the traditional definition [Santos97]:

$$S = \frac{1}{N_{row}N_{col}} \sum_{x=1}^{N_{row}} \sum_{y=1}^{N_{col}} \left[f(x,y) - \bar{f} \right]^2 . \tag{2.8}$$

Several sharpness measures are derived based on the image histogram. The most straightforward measure is the difference between the maximum and minimum grey levels [Santos97]. Another popular choice uses the entropy of the image grey levels [Santos97]. Kroktov also proposed a measure using the histogram of local variations [Kroktov89].

2.3.3 Autocorrelation based measures

Autocorrelation evaluates the dependency among neighboring pixels, which provides another practical way to quantify image sharpness. In literature, some of the sharpness measures simply compute one sample of the autocorrelation function as given by [Kroktov89]:

$$S = \sum_{x=1}^{N_{row}-1} \sum_{y=1}^{N_{col}} f(x,y) f(x+1,y) - \sum_{x=1}^{N_{row}-2} \sum_{y=1}^{N_{col}} f(x,y) f(x+2,y) . \tag{2.9}$$

More complicated measures use quantities such as the area [Batten00] and the height [Ong98] of the central peak of the autocorrelation function.

2.3.4 Transform based measures

In this category, the image is first transformed into the frequency domain usually via Fourier transform (FT) or discrete cosine transform (DCT). The sharpness measure is then computed based on the coefficients $F(u,v)$ in the frequency domain or their distributions. The Fast Fourier Transform (FFT) sharpness measure is defined as [Subbarao92]:

$$S = \sum\sum |\, Magnitude(u,v) \times Angle(u,v) \,| \,. \tag{2.10}$$

The sum of the amplitudes of the frequency coefficients within a predefined window W_F

$$S = \sum_{(u,v)\in W_F} \sum |\, F(u,v) \,|\,, \tag{2.11}$$

is also used as sharpness measure [Batten00].

Besides point based definitions, some measures explore the statistical information contained in the frequency domain. The multivariate kurtosis, derived from the distribution of the FT coefficients, is employed as a sharpness metric [Zhang99]. Kristan *et al.* proved that the maximum entropy in the frequency domain coincides with the maximum sharpness in the spatial domain and proposed an entropy based measure [Kristan04].

2.3.5 Edge based measures

Edge based measures make use of the edge components, which are primarily responsible for the visual perception of image sharpness. In theory, edge based methods should better represent the sharpness of an image. However, these approaches are not widely used mainly because of the computational complexities associated with edge detection and characterization.

Li defined an ideal 2D step edge as [Li02]:

$$f(x,y) = f_o(x,y) + \frac{c}{2}\left[1 + erf\left(\frac{x\cos\theta + y\sin\theta}{\sqrt{2}w}\right)\right], \tag{2.12}$$

where c, θ, and w represent the contrast, orientation and scale, respectively, $f_o(x,y)$ is the mean intensity level, and $erf(\circ)$ denotes the error function. The scale w describes the width of the edge transition, whose average value determines image sharpness. The proposed algorithm provides a neat solution in theory. However, it requires the isolation of step edges. A filter bank, adjusted to various edge orientations, was used by Dijk *et al.* to detect the average edge width [Dijk02].

As an improvement over the global kurtosis sharpness measure [Zhang99], Caviedes and Gurbuz proposed a local kurtosis sharpness measure based on both spatial edges and

coefficients in the transformed domain [Caviedes02]. Compared with other edge based algorithms, the local kurtosis measure handles different types of edges in the same fashion and avoids the difficulty in distinguishing step and line edges.

2.3.6 Performance comparison

According to our survey, the characteristic behaviors of various sharpness measures can be summarized as follows.

(1) Gradient based measures yield a performance closest to the ideal response and more importantly their performances are robust to degradations introduced by high magnification. However, their values drop and saturate rapidly as the focus moves away from the optimal position, resulting in a large portion of flat response regardless of the changes in the camera's focus. Given an initial focus position in the saturation region, it is difficult to determine the direction that leads to an increased sharpness value.

(2) Statistics based measures perform global operations, such as computing image variance and histogram, and neglect the local information of image edges, which is responsible for their inferior accuracy.

(3) The performance of the autocorrelation based measures is comparable to that of the gradient based measures. In addition, the decreasing/increasing slope is adjustable by choosing different window sizes. With a smaller window size, the response is relatively sharp and narrow similar to that of the gradient based measures, while measures with a larger window size produce wide peaks and gradual slopes. This feature can be used to balance two criteria during focus search: precise location and easy direction initialization.

(4) As to the measures defined in the transformed domain, their performance falls in between the gradient based and statistics based measures. The associated computations depend on the transform used.

(5) Edge based measures yield comparable performances as transform based measures. However, their computational complexity is substantially intensive. In addition, it is difficult to detect strong edges in a blurred high magnification image. Thus its applicability to high magnification images remains questionable. Table 2.2 summarizes the comparison.

Table 2.2. Comparison of sharpness measures.

Measures	Advantages	Disadvantages
Gradient based	Sharp peak Low computational complexity	Large portion of saturation region
Statistics based	Low computational complexity	Low accuracy and noisy response
Autocorrelation based	Response slope is adjustable	Slightly increased computations
Transform based	Sharp peak	Slightly increased computations
Edge based	Representative of visual perception	High computational complexity Difficulties in separating strong edges

3 Sensor planning

With the increased scale and complexity involved in most practical surveillance applications, it is almost impossible for any single camera (either omnidirectional or PTZ) to fulfill tracking and monitoring with an acceptable degree of continuity and/or a reasonable accuracy. Systems with multiple cameras enter into play and find extensive applications. The concept of sensor planning comes naturally when the question of how to place multiple cameras for the best coverage and at the lowest cost arises.

A descriptive definition of sensor planning given in [Tarabanis95] is quoted as: "Given information about the environment as well as the information about the task that the vision system is to accomplish, develop strategies to automatically determine sensor parameter values that achieve this task with a certain degree of satisfaction." When formulated mathematically as an optimization process, there exist two types of problems in sensor planning: (1) the search for the maximum coverage given a fixed total cost or number of cameras and (2) the search for the minimum cost or number of cameras for a full or required coverage [Erdem06, Lee91]. In this paper, we refer to (1) and (2) as the Max-Coverage (Type 1) and Min-Cost (Type 2) problems.

Assuming that a polygonal floor plan is represented as an occupancy grid, a binary vector $\mathbf{b}$ can be obtained by letting $b_i = 1$ if the i^{th} grid can be seen by at least one camera and $b_i = 0$ otherwise. We construct a binary matrix A with $a_{ij} = 1$ if the i^{th} grid is covered by the j^{th} camera configuration. Each camera configuration specifies one combination of the camera's intrinsic and extrinsic parameters, including the camera's focal length f, pan/tilt angle θ_P / θ_T, and position T_C. The following relation holds: $b_i = 1$ if $b_i' > 0$ and $b_i = 0$ otherwise, with $\mathbf{b}' = A\mathbf{x}$. The solution vector $\mathbf{x}$ specifies a set of chosen camera configurations with the corresponding element x_j=1 if the configuration is chosen and x_j=0 otherwise.

Let the cost associated with the j^{th} camera configuration be ω_j. Given the maximum cost $C_{\max}$, the Max-Coverage problem can be described by:

$$\max \sum_i b_i, \text{ subject to } \sum_j \omega_j x_j \le C_{\max} \quad . \tag{3.1}$$

Given a specified coverage vector $\mathbf{b}_{C,o}$ or a minimum overall coverage $C_{\min}$, the Min-Cost problem can be modeled as:

$$\min \sum_j \omega_j x_j, \text{ subject to } A\mathbf{x} \ge \mathbf{b}_{C,o} \text{ or } \sum_i b_i \ge C_{\min} . \tag{3.2}$$

In addition to the conventional requirements in sensor planning, such as coverage and cost, extra criteria need to be considered to ensure persistent tracking and monitoring in a real-time automatic surveillance system. One of the criteria to be included is a sufficient amount of overlapped FOVs between adjacent cameras so that enough time is reserved to perform consistent labeling and camera handoff. This criterion, to which this chapter is devoted, is however not addressed in existing camera placement algorithms. Thereby, our algorithm improves existing camera placement methods by adding handoff rate analysis.

In coverage analysis, only two types of areas, visible and invisible, are used. To incorporate handoff rate analysis, a third type of area, handoff safety margin, is introduced, which defines visible areas requiring camera handoff. An observation measure is proposed to define the handoff safety margin. We then develop sensor planning algorithms balancing the tradeoff between overall coverage and adequate overlapped handoff safety margins. Variations, such as direct constraint and adaptive weight approaches, are introduced for special considerations of resolution and frontal view. Furthermore, the problems of dynamic occlusion and camera overload are addressed so that the optimal handoff success rate can be achieved regardless of the dynamic interactions among multiple moving targets and the camera's limited computational capacities.

Figure 3.1 illustrates the flow chart of our sensor planning algorithm. In parallel with the definition of sensor planning given in [Tarabanis95], our algorithm has three inputs: environment representation, camera modeling, and performance requirements. Tracking and observation suitability is evaluated via the observation measure and thresholds separating the visible area, handoff safety margin, and invisible area are obtained from target behavior modeling. Based on these three areas an objective function is constructed and used to guide the search for the optimal camera placement.

The remainder of this chapter is organized as follows. Section 3.1 defines the observation measure. The objective function is described in section 3.2 with experimental results demonstrated in section 3.3.

3.1 Observation measure

In addition to visibility, we introduce the following criteria to describe the

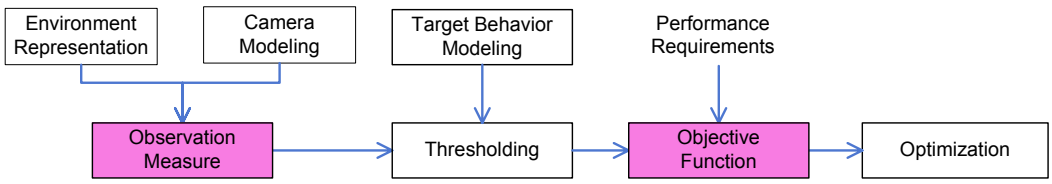


Figure 3.1. Flow chart of sensor planning.

observation of the tracked target: its resolution M_R, its distance to the edges of the camera's FOV M_D, and the availability of a frontal view M_{FV}. From a viewer's perspective, visibility is the fundamental requirement. Herewith, the viewer includes not only human operators but also successive automatic processing such as consistent labeling, object tracking, and face/object recognition. Observations with different detail levels affect the performance of these algorithms. For example, a frontal face image with an inter-ocular distance no smaller than 60 pixels is recommended by a well-known face recognition engine FaceIt® for a face to be automatically recognized [Phillips02]. For persistent object tracking and smooth camera handoff, the tracked target should be at a reasonable distance from the edges of the camera's FOV. The M_D component considers the margin for executing handoff before the object falls out of the camera's FOV.

3.1.1 Static perspective cameras

To begin our study, the camera and world coordinates are defined and illustrated in Figure 3.2. A point $[X \quad Y \quad Z]^T$ in the world coordinates is projected onto a point $[x' \quad y' \quad z']^T$ in the camera coordinates by:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos\theta_T & 0 & -\sin\theta_T \\ 0 & 1 & 0 \\ \sin\theta_T & 0 & \cos\theta_T \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_P & \sin\theta_P \\ 0 & -\sin\theta_P & \cos\theta_P \end{bmatrix} \begin{bmatrix} Z - T_Z \\ X - T_X \\ Y - T_Y \end{bmatrix}, \tag{3.3}$$

with $T_C = [T_X \quad T_Y \quad T_Z]^T$. Assuming zero skew, unit aspect ratio, and image center on the principal point, the projected point in the image plane is given by: $\begin{cases} x = fx'/z' \\ y = fy'/z' \end{cases}$. Letting $Z = 0$ (points on the ground plane), we have:

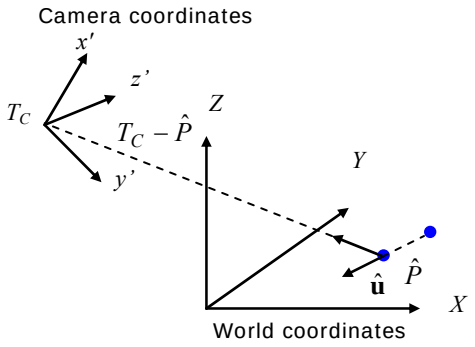


Figure 3.2. Illustration of the camera and world coordinates for perspective cameras.

$$\begin{cases} x = f \dfrac{-T_Z \cos\theta_T - Z' \sin\theta_T}{-T_Z \sin\theta_T + Z' \cos\theta_T} \\ y = f \dfrac{Y'}{-T_Z \sin\theta_T + Z' \cos\theta_T} \end{cases}, \tag{3.4}$$

where

$$\begin{bmatrix} Y' \\ Z' \end{bmatrix} = \begin{bmatrix} \cos\theta_P & \sin\theta_P \\ -\sin\theta_P & \cos\theta_P \end{bmatrix} \begin{bmatrix} X - T_X \\ Y - T_Y \end{bmatrix}. \tag{3.5}$$

The estimation of the target depth $\hat{z}'$ can be obtained by:

$$\hat{z}' = -T_Z \sin\theta_T + Z' \cos\theta_T = \frac{-T_Z}{x / f \cos\theta_T + \sin\theta_T}. \tag{3.6}$$

For static cameras with a constant focal length, the estimated target depth is sufficient to describe the resolution:

$$M_R = \alpha_R / \hat{z}', \tag{3.7}$$

where α_R is a normalization coefficient. However, when the target is at a close distance, this relation is not entirely valid, especially when part of the target falls out of the camera's FOV. Therefore, the above definition is modified:

$$M_R = \begin{cases} \alpha_R / \hat{z}' & \hat{z}' > -T_Z / \tan\theta_T \\ \dfrac{\alpha_R}{\left(\hat{z}' + T_Z / \tan\theta_T\right)^2 - T_Z / \tan\theta_T} & \hat{z}' \le -T_Z / \tan\theta_T \end{cases}. \tag{3.8}$$

In practice, for a better observation and to reserve enough computation time for camera handoff, the target should remain at a safe distance from the edges of the camera's FOV. Moreover, this margin distance is affected by the target depth. When the target is at a closer distance, its projected image undergoes larger displacements in the image plane. Therefore, a larger margin should be reserved. In our definition, different polynomial powers are used to achieve varying decreasing/increasing rates. The M_D is then given by:

$$M_D = \left\{ \alpha_D \left[\left(1 - \frac{|x|}{N_{row}/2}\right)^2 + \left(1 - \frac{|y|}{N_{col}/2}\right)^2 \right] \right\}^{\beta_1 \hat{z}' + \beta_0}, \tag{3.9}$$

where N_{col} and N_{row} denote the image's width and height, α_D is a normalization weight, and coefficients β_1 and β_0 are used to adjust the polynomial power.

The frontal view measure computes the angle between the target's motion direction $\hat{\mathbf{u}}$ and the direction of the line connecting the target's current position $\hat{P}$ and the camera's optical center:

$$M_{FV} = \frac{\alpha_{FV}\left(T_C - \hat{P}\right)^T \hat{\mathbf{u}}}{\| T_C - \hat{P} \|\| \hat{\mathbf{u}} \|}, \tag{3.10}$$

where α_{FV} is a normalization factor. The observation measure for a static perspective camera is then given by:

$$Q = \begin{cases} w_R M_R + w_D M_D + w_{FV} M_{FV} & [x\ y]^T \in \Pi \\ -\infty & otherwise \end{cases}, \tag{3.11}$$

where w_R, w_D, and w_{FV} are importance weights and Π denotes the image plane.

3.1.2 PTZ cameras

For PTZ cameras with varying zooms, the resolution component M_R is given by:

$$M_R = \alpha_R\, f / \hat{z}'. \tag{3.12}$$

Compared with (3.8), the additional term for the special case when part of the target falls out of the camera's FOV is not necessary because of the additional flexibility from the camera's adjustable tilt angle. In addition, we assume that the target is always maintained at the image center by panning and tilting the camera. Therefore, the M_D component can be eliminated from the computation of the observation measure.

However, the assumption that the target is always maintained at the image center sometimes requires extreme pan and tilt speeds. Let the instant FOV denote the FOV that a PTZ camera can see at any given time instance and the achievable FOV the FOV that a PTZ camera can survey given a sufficient period of time. The limited pan and tilt speeds lead to the discrepancy between the instant FOV and the achievable FOV. To address this issue, a common practice is to impose additional constraints on the maximum time duration for a PTZ camera to pan and tilt to a specified position. We will come back to this issue in the discussion of our sensor planning algorithm for multiple dynamic targets, where the aforementioned discrepancy is resolved elegantly using the probability of camera overload.

The definition of the frontal view component remains the same and the observation measure is a weighted sum of the M_R and M_{FV} components.

3.1.3 Omnidirectional cameras

The geometry of an omnidirectional camera is depicted in Figure 3.3. The imaging process of an omnidirectional camera does not comply with the traditional perspective projection. Let r denote the distance between the projected point $[x\ y]^T$ and the principal point and θ the angle between the incoming ray and the optical axis. The perspective projection is characterized by $r = f\tan\theta$. To realize a wider opening angle, this relation is changed. Various projection models exist in literature [Kannala04], such as the equidistance projection $r = f\theta$ and the general polynomial model $r = f\sum_{k=1,odd}\lambda_{\theta,k}\theta^k$ where $\lambda_{\theta,k}$ denote the approximation coefficients. Image resolution is the partial derivative of r with respect to R:

$$M_R = \alpha_R \frac{\partial r}{\partial R} = \frac{\alpha_R fZ}{Z^2 + R^2} \sum_{k=1,odd} \lambda_{\theta,k} k\theta^{k-1}, \tag{3.13}$$

with $R = \sqrt{X^2 + Y^2}$. The M_D component is given by:

$$M_D = \alpha_D (1 - r / r_o)^2, \tag{3.14}$$

where r_o represents the image size of the omnidirectional camera. The definition of the frontal view component remains the same and the observation measure is a weighted sum of the M_R, M_D, and M_{FV} components.

3.1.4 Handoff safety margin

A failure threshold Q_F and a trigger threshold Q_T are derived to define three disjoint regions: (1) invisible area with $Q_{ij} < Q_F$ where Q_{ij} represents the observation measure value of the i^{th} grid observed by the j^{th} camera configuration, (2) visible area with $Q_{ij} \geq Q_T$,

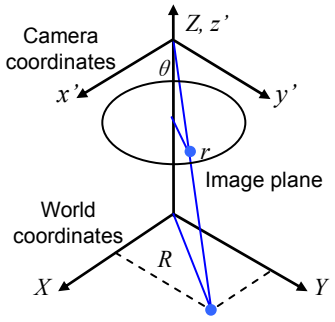


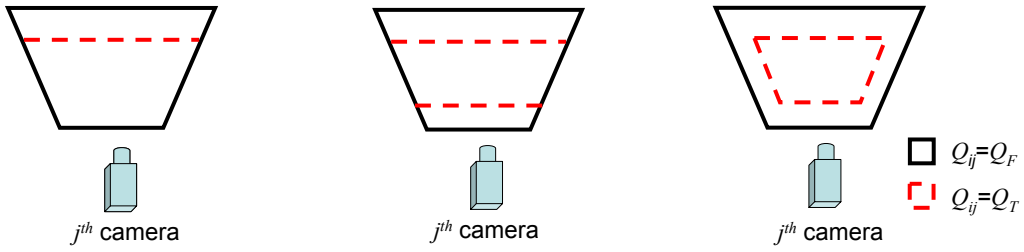
Figure 3.3. Illustration of the geometry for omnidirectional cameras.

and (3) handoff safety margin with $Q_F \le Q_{ij} < Q_T$. The failure threshold Q_F segments the invisible areas and is used for coverage analysis. The trigger threshold Q_T separates the visible areas and handoff safety margins. It is introduced for handoff rate analysis, where necessary overlapped FOVs between adjacent cameras are optimized. The trigger threshold Q_T is given by $Q_T = Q_F + \kappa u_{obj} t_H$ where u_{obj} represents the average moving speed of the object of interest, t_H denotes the average duration for a successful handoff, and κ is a conversion scalar.

The individual and combined effects of the M_R and M_D components become evident when we study the contours of the observation measure defined by Q_F and Q_T. In Figure 3.4, the black solid lines and red dashed lines depict the contours with $Q_{ij} = Q_F$ and $Q_{ij} = Q_T$, respectively. The resolution component M_R provides limits along the direction of the camera's optical axis while the M_D component generates constraints mainly in the direction orthogonal to the camera's optical axis. If (3.7) is used as shown in Figure 3.4(a), the handoff safety margin is given by $\alpha_R / \hat{z}' < Q_T$. That is $\alpha_R / Q_T < \hat{z}'$. As a result, the handoff safety margin is only defined at the far end of the camera's FOV along the optical axis. The scenario where the target is so close to the camera that part of it falls out of the camera's FOV is ignored. The modification in (3.8) imposes a proper constraint at the near end of the camera's FOV along the optical axis, as shown in Figure 3.4(b). Therefore, the resulting observation is complete and with the desired resolution.

3.2 Objective function

Let A_1 represent the grid coverage with $a_{1,ij} = 1$ if $Q_{ij} \ge Q_F$ and $a_{1,ij} = 0$ otherwise. The A_1 matrix resembles the A matrix in the conventional coverage analysis discussed in the previous section. Two additional matrices are constructed A_2 and A_3. The matrix A_2 has



(a) (b) (c)

Figure 3.4. Schematic illustration of the contours of the observation measure with $Q_{ij} = Q_F$ and $Q_{ij} = Q_T$ to show the effect of the M_R and M_D components. (a) $Q = M_R = \alpha_R / \hat{z}'$. (b) $Q = M_R$ as defined in (3.8). (c) $Q = w_R M_R + w_D M_D$ with $w_R = 0.5$ and $w_D = 0.5$.

$a_{2,ij}=1$ if $Q_F \le Q_{ij} < Q_T$ and $a_{2,ij}=0$ otherwise. The matrix A_3 has $a_{3,ij}=1$ if $Q_{ij} \ge Q_T$ and $a_{3,ij}=0$ otherwise. Matrices A_2 and A_3 represent the handoff safety margin and visible area, respectively. Let $\mathbf{c'}_k = A_k \mathbf{x}$, $k=1,2,3$. The objective function is formulated as:

$$c_i = w_1(c'_{1,i}>0) + w_2(c'_{2,i}=2) - w_3(c'_{3,i}>1), \tag{3.15}$$

where w_1, w_2, and w_3 are predefined importance weights. The operation $(c'_{1,i}>0)$ means $(c'_{1,i}>0)=\begin{cases}1 & c'_{1,i}>0\\ 0 & otherwize\end{cases}$. The first term in the objective function considers coverage, the second term produces sufficient overlapped handoff safety margins, and the third term penalizes excessive overlapped visible areas. Our objective function achieves a balance between coverage and sufficient margins for camera handoff. The optimal sensor placement for the Max-Coverage and Min-Cost problems can then be obtained by:

$$\max \sum_i c_i, \text{ subject to } \sum_j \omega_j x_j \le C_{\max} \quad , \tag{3.16}$$

$$\min \sum_j \omega_j x_j \text{ then } \max \sum_i c_i, \text{ subject to } \quad A\mathbf{x} \ge \mathbf{b}_{C,o} \text{ or } \sum_i b_i \ge C_{\min}. \tag{3.17}$$

3.2.1 Function validation

To validate our objective function, we consider the positioning of two cameras for example. Figure 3.5 shows the relative position of two perspective cameras, where the FOV of camera 1 is centered at the origin of the world coordinates in the ground plane and camera 2 is free to translate $(\Delta X, \Delta Y)$ and rotate $(\Delta\theta_P, \Delta\theta_T)$. From the definition of the observation measure, the contours defined by $Q_{ij}=Q_F$ and $Q_{ij}=Q_T$ approximate trapezoids. The corresponding parameters are given in Figure 3.5.

The derivation of the exact expression of the objective function is not difficult but tedious. To simplify the process and yet reveal the characteristics of the objective function, we fix $\Delta Y = \Delta\theta_P = \Delta\theta_T = 0$ and study the relation between the objective function $F=\sum_i c_i$ and ΔX as our first step. The resulting function can be expressed as:

$$\begin{cases} F_1 = 2w_1(L_F + l_F)h_F + (w_2 - w_1)\dfrac{(\Delta X - 2L_F)^2 h_F}{4(L_F - l_F)} & \Delta X_{th} < \Delta X \le 2L_F \\ F_2 = F_1 - w_2 \dfrac{(\Delta X - \Delta X_{th})^2 h_F}{4(L_F - l_F)} & 2\tau L_F < \Delta X \le \Delta X_{th} \\ F_3 = F_2 - w_3 \dfrac{(\Delta X - 2\tau L_F)^2 h_F}{4(L_F - l_F)} & \tau L_F \le \Delta X \le 2\tau L_F \end{cases}, \tag{3.18}$$

with $\Delta X_{th} = 2L_F - (1-\tau)(3L_F - l_F)/2$ and $\tau = L_T / L_F$. Since the coverage, overlapped handoff margins, and overlapped visible areas become effective in (3.15) in sequence as ΔX decreases, F has three expressions depending on the value of ΔX. Given the

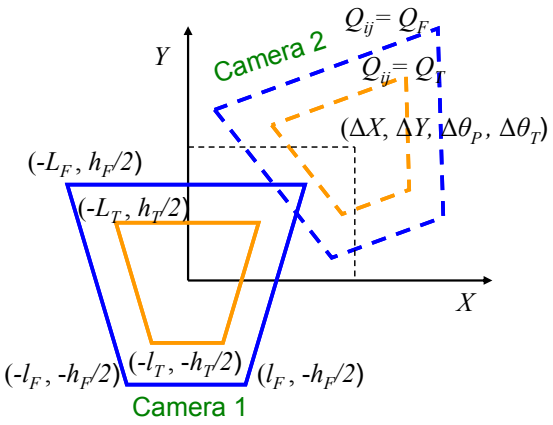


Figure 3.5. Schematic illustration of the geometry relation between the adjacent cameras' FOVs for the computing of the objective function. The position of camera 1 is fixed while camera 2 is free to translate and rotate. Both cameras are static perspective cameras.

expression in (3.18), Figure 3.6 shows the objective function with different choices of weights. We can see that the optimal ΔX^* is achieved with $2\pi L_F < \Delta X^* \leq \Delta X_{th}$. When a smaller weight is assigned to the coverage term, the optimal ΔX^* is shifted toward $2\pi L_F$, resulting in more overlapped FOVs for executing camera handoff. From the derivatives of (3.18), we note that F_1 is a monotonously decreasing function if $w_2 > w_1$. With a proper choice of w_3, F_3 is a monotonously increasing function. As a result, the turning point falls in the range of F_2 and is determined by the relation between w_1 and w_2, the weights for the coverage and handoff margin terms. Figure 3.7 shows the objective function as a function of ΔX and ΔY.

Since the observation measures for omnidirectional cameras are radial symmetric, it is sufficient to study the variations along the radial direction. Figure 3.8 shows the FOVs of two omnidirectional cameras placed ΔR distance apart. We want to examine the behavior of our objective function with varying ΔR. The contours defined by $Q_{ij} = Q_F$ and $Q_{ij} = Q_T$ are concentric circles with radii of R_F and R_T, respectively.

Figure 3.9 depicts the values of the objective function $F = \sum_i c_i$ as a function of ΔR. Different choices of w_1 are used to illustrate their influence on the optimal camera position. The optimal camera position is achieved with $2R_T \leq \Delta R^* \leq R_F + R_T$. The actual position depends on the w_1 used. Like the case of perspective cameras, a smaller w_1 results in a camera placement with a smaller ΔR^*. The exact expression and derivative of the objective function for omnidirectional cameras are given in (3.19) and (3.20), respectively.

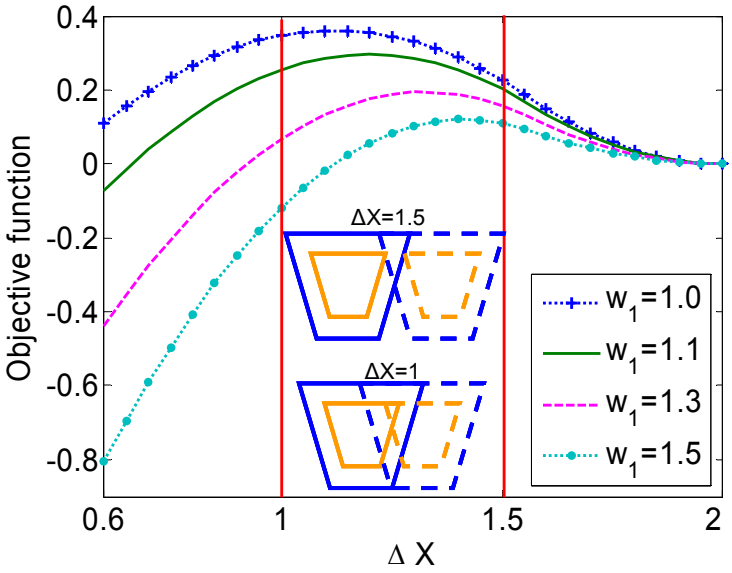


Figure 3.6. The objective function for perspective cameras with varying ΔX and different choices of w_1, the weight assigned to the coverage term in (3.15). w_2=2, w_3=5, L_F=1, l_F=0.6, h_F=0.8, τ=0.6.

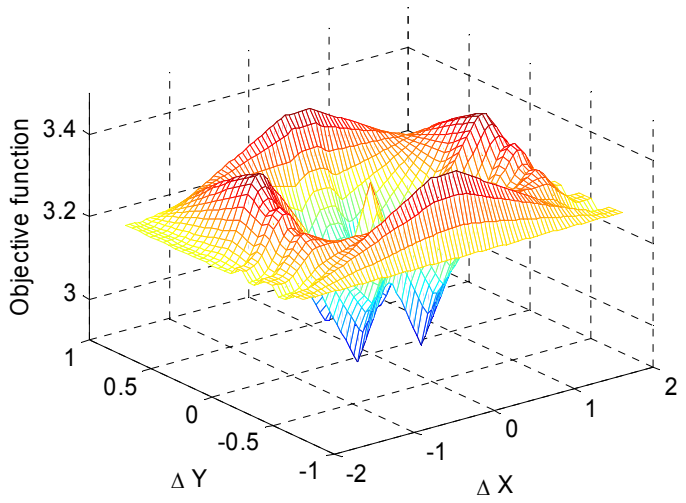


Figure 3.7. The objective function for perspective cameras with varying ΔX and ΔY. The weights are w_1=1.2, w_2=2, and w_3=5. L_F=1, l_F=0.6, h_F=0.8, τ=0.6.

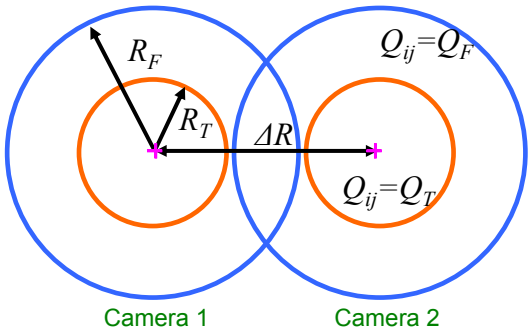


Figure 3.8. Illustration of the FOVs in the ground plane (Z=0) of two omnidirectional cameras. The position of camera 1 is fixed while camera 2 is free to translate.

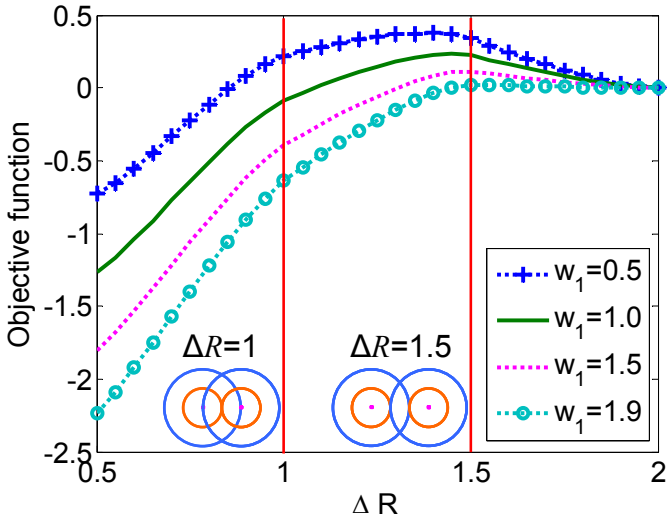


Figure 3.9. The objective function for omnidirectional cameras with varying ΔR and different choices of w_1, the weight assigned to the coverage term in (3.15). w_2=2, w_3=5, R_F=1, R_T=0.5.

$$\begin{cases} F_1 = w_1 \pi R_F^2 + (w_2 - w_1)\left[R_F^2 \cos^{-1}\left(\dfrac{\Delta R}{2R_F}\right) - \dfrac{\Delta R}{2}\sqrt{R_F^2 - \left(\dfrac{\Delta R}{2}\right)^2} \right] & R_F + R_T \le \Delta R \le 2R_F \\ F_2 = F_1 - w_2\left[R_T^2 \cos^{-1}\left(\dfrac{\Delta R^2 - R_T^2 + R_F^2}{2\Delta R R_T}\right) + R_F^2 \cos^{-1}\left(\dfrac{\Delta R^2 - R_F^2 + R_T^2}{2\Delta R R_F}\right) \right. & \\ \qquad \left. - \dfrac{1}{2}\sqrt{(-\Delta R + R_F + R_T)(\Delta R + R_F - R_T)(\Delta R - R_F + R_T)(\Delta R + R_F + R_T)} \right] & 2R_T \le \Delta R < R_F + R_T \\ F_3 = F_2 - w_3\left[R_T^2 \cos^{-1}\left(\dfrac{\Delta R}{2R_T}\right) - \dfrac{\Delta R}{2}\sqrt{R_T^2 - \left(\dfrac{\Delta R}{2}\right)^2} \right] & R_T \le \Delta R < 2R_T \end{cases} \tag{3.19}$$

$$\begin{cases} F_1' = -(w_2 - w_1)\sqrt{R_F^2 - \left(\dfrac{\Delta R}{2}\right)^2} & R_F + R_T \le \Delta R \le 2R_F \\ F_2' = F_1' + w_2\sqrt{\dfrac{-\Delta R^4 - R_F^4 - R_T^4 + 2\Delta R^2 R_F^2 + 2\Delta R^2 R_T^2 + 2R_F^2 R_T^2}{\Delta R^2}} & 2R_T \le \Delta R < R_F + R_T \\ F_3' = F_2' + w_3\sqrt{R_T^2 - \left(\dfrac{\Delta R}{2}\right)^2} & R_T \le \Delta R < 2R_T \end{cases} \tag{3.20}$$

Finally, Figure 3.10 presents a plot of the objective function for omnidirectional cameras as a function of ΔX and ΔY. We could see that the maxima of the objective function are obtained at $\sqrt{\Delta X^2 + \Delta Y^2} = \Delta R^*$ due to the radial symmetric property of the omnidirectional cameras.

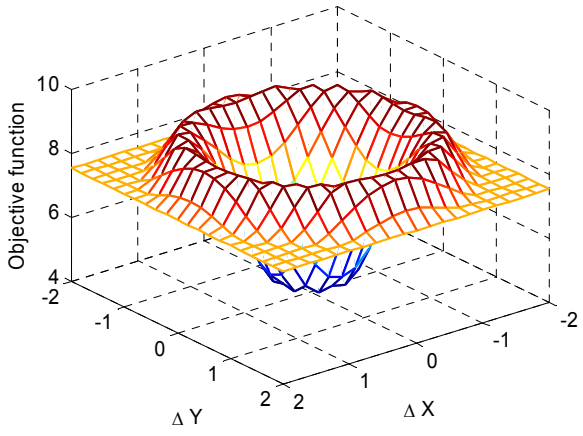


Figure 3.10. The objective function for omnidirectional cameras with varying ΔX and ΔY. The weights are w_1=1.2, w_2=2, and w_3=5. R_F=1, R_T=0.5.

3.2.2 Environments with multiple dynamic targets

Environments with multiple moving objects impose additional difficulties on sensor planning. Multiple moving objects cause dynamic occlusions depending on their real-time relative positions. Figure 3.11 compares two camera placements in terms of the ability to handle dynamic occlusion. It is obvious that the camera placement in Figure 3.11(a) is unable to deal with dynamic occlusion since target 2 is blocked by target 1 in the FOVs of both cameras. On the contrary, in the camera placement shown in Figure 3.11(b), target 2 can be seen from camera 2 when it is occluded by target 1 in camera 1. From the above illustration, we could see that the probability of dynamic occlusion can be reduced by a proper camera placement. Due to the non-deterministic nature of dynamic occlusion, analysis regarding such occlusions is conducted in a probabilistic framework. The probability of dynamic occlusion P_{do} is derived and incorporated into sensor planning.

Another important issue in sensor planning for environments with multiple dynamic targets is the coordination among multiple cameras. In practice, a single camera can track a limited number of targets simultaneously because of the limited resolvable distance and computational capacities. The camera may not be able to detect and/or track new objects when its maximum computational capacity has been reached. This scenario is referred to as the problem of camera overload and is demonstrated in Figure 3.12. Assume that the camera is able to track four targets at maximum simultaneously. When a new target enters the camera's FOV, a decision is to be made so that an appropriate target is dropped due to the limited computational capacity. In Figure 3.12(b), since target 3 is farther away from the camera, it is dropped so that the camera can track the new target. The goal of sensor planning is to automatically minimize the number of dropped targets due to camera overload.

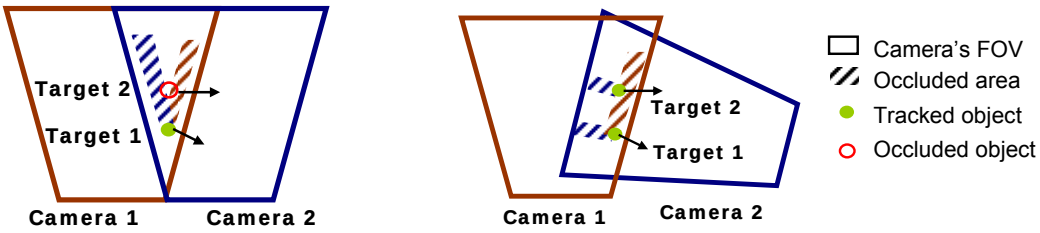
For dynamic occlusion analysis, we follow the approach proposed by Mittal and Davis [Mittal04]. Objects are modeled as a cylinder with a radius of r_{obj} and a height of h_{obj}. Let the area of their projection onto the ground plane be fixed as $A_{ob} = \pi r_{obj}^2$. Assume that the object of interest centered at the i^{th} grid is observed by the j^{th} camera from a distance D_{ij}, as shown in Figure 3.13. Its region of occlusion is:

$$A_{o,ij} = 2 r_{obj} h_{obj} \frac{D_{ij}}{T_{Z,j}}. \tag{3.21}$$

Assuming a uniform object density, the occlusion probability at the i^{th} grid observed by the j^{th} camera $P_{do,ij}$ can then be expressed as [Mittal04]:

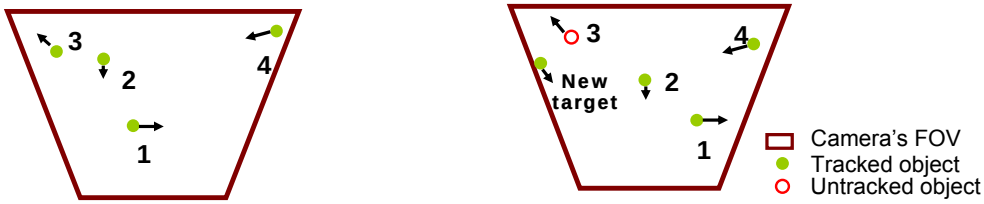
$$P_{do,ij} = 1 - \lim_{K \to \infty} \prod_{k=0}^{K-1} \left(1 - \frac{A_{o,ij}}{K / K_o - k A_{ob}} \right), \tag{3.22}$$

where K_o denotes the object density. Under the assumption that K_o is much smaller than $1/A_{ob}$, $P_{do,ij}$ can be further simplified:



(a) (b)

Figure 3.11. Schematic illustration of the problem of dynamic occlusion. (a) Target 2 is occluded by target 1 in both cameras. (b) Target 2 can be observed from camera 2 when it is occluded by target 1 in camera 1.



(a) (b)

Figure 3.12. Schematic illustration of the problem of camera overload. Assume that the camera is able to track four targets at maximum simultaneously due to limited computational capacities. (a) The maximum number of targets is achieved. (b) Camera overload occurs when a new target enters the camera's FOV. Target 3 is dropped due to camera overload.

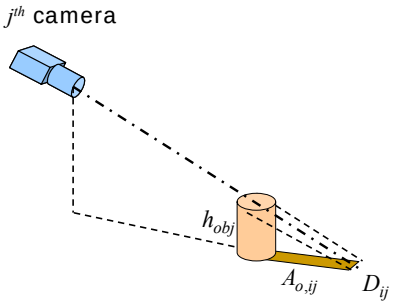


Figure 3.13. Illustration of the region of occlusion.

$$P_{do,ij} = 1 - \exp\left\{ -\frac{K_o A_{o,ij} (2 - K_o A_{ob})}{2(1 - K_o A_{ob})} \right\}. \tag{3.23}$$

The overall probability of dynamic occlusion at the i^{th} grid $P_{do,i}$ is:

$$P_{do,i} = \prod_{j, a_{1,ij} x_j = 1} P_{do,ij} \ . \tag{3.24}$$

The objective function becomes:

$$c_i = w_1(c'_{1,i} > 0) + w_2(c'_{2,i} = 2) - w_3(c'_{3,i} > 1) + w_4(P_{do,i} \le P_{do,th}), \tag{3.25}$$

where $P_{do,th}$ is a predefined threshold and w_4 is the importance weight.

For camera overload analysis, we consider the multi-object tracking system as an $\mathcal{M}/\mathcal{M}/\mathcal{N}/\mathcal{N}$ queuing system. Following the conventions in queuing theory, an $\mathcal{M}/\mathcal{M}/\mathcal{N}/\mathcal{N}$ system suggests that: (1) the arrival process follows a Poisson distribution; (2) the residence time follows an exponential distribution; and (3) the number of servers and buffer slots are $\mathcal{N}$. Denote the average arrival rate in the FOV of the j^{th} camera as $\lambda_{c,j}$ and the mean camera-residence time as $1/\mu_{c,j}$. Let $N_{obj,j}$ be the maximum number of targets that can be tracked simultaneously by the j^{th} camera. From the $\mathcal{M}/\mathcal{M}/\mathcal{N}/\mathcal{N}$ queuing theory, the system can be described by a Markov chain. Given the probability of the (n-1)th state $P_{n-1,j}$, the probability of the n^{th} state $P_{n,j}$ is expressed as:

$$P_{n,j} = \frac{\lambda_{c,j}}{n\mu_{c,j}} P_{n-1,j} = \frac{1}{n!}\left(\frac{\lambda_{c,j}}{\mu_{c,j}}\right)^n p_{o,j} \ , \tag{3.26}$$

for $1 \le n \le N_{obj,j}$ where $p_{o,j}$ is a normalization term to make the sum of the probabilities of all possible states as one.

$$\sum_{n=0}^{N_{obj,j}} P_{n,j} = \sum_{n=0}^{N_{obj,j}} \left[\frac{1}{n!}\left(\frac{\lambda_{c,j}}{\mu_{c,j}}\right)^n p_{o,j} \right] = 1 \Rightarrow p_{o,j} = \left[\sum_{n=0}^{N_{obj,j}} \frac{1}{n!}\left(\frac{\lambda_{c,j}}{\mu_{c,j}}\right)^n \right]^{-1} \tag{3.27}$$

The probability that the j^{th} camera reaches its maximum computational capacity is the probability that the Markov chain reaches the ($N_{obj,j}$)th state:

$$P_{\max,j} = \frac{\frac{1}{N_{obj,j}!}\left(\frac{\lambda_{c,j}}{\mu_{c,j}}\right)^{N_{obj,j}}}{\sum_{n=0}^{N_{obj,j}} \frac{1}{n!}\left(\frac{\lambda_{c,j}}{\mu_{c,j}}\right)^n} . \tag{3.28}$$

Denote the average arrival rate at the i^{th} grid as $\lambda_{g,i}$ and the mean camera-residence time as $1/\mu_{g,i}$. The probability of camera overload at the i^{th} grid $P_{co,i}$ is given by:

$$P_{co,i} = (1 - e^{-\lambda_{g,i}/\mu_{g,i}}) \prod_{j, a_{1,ij} x_j = 1} P_{\max,j} \ . \tag{3.29}$$

The objective function becomes:

$$c_i = w_1(c'_{1,i} > 0) + w_2(c'_{2,i} = 2) - w_3(c'_{3,i} > 1) + w_5(P_{co,i} \le P_{co,th}), \tag{3.30}$$

where $P_{co,th}$ is a predefined threshold and w_5 denotes the importance weight.

The significance of introducing camera overload analysis becomes obvious especially for PTZ cameras. As mentioned in section 3.1.2, camera placement algorithms always find it difficult to properly model the PTZ camera's instant and achievable FOVs. At a given time instance, a PTZ camera has a limited instant FOV. However, given enough time to pan and tilt, a PTZ camera has a 360°×90° achievable FOV. The discrepancy in modeling the PTZ camera's instant and achievable FOVs is solved by letting $N_{obj} = 1$. The limited instant FOV can be described as the achievable FOV with $N_{obj} = 1$. That is at a given time instance, a single PTZ camera is able to track a single object in its 360°×90° achievable FOV. In Figure 3.14, target 1 is tracked by the PTZ camera from t_{k_o} to t_{k_1}. If another object enters the camera's 360°×90° achievable FOV, it cannot be seen since the camera's instant FOV points to the current tracked object. This agrees with the reasoning based on the achievable FOV with $N_{obj} = 1$. As the maximum number of tracked objects

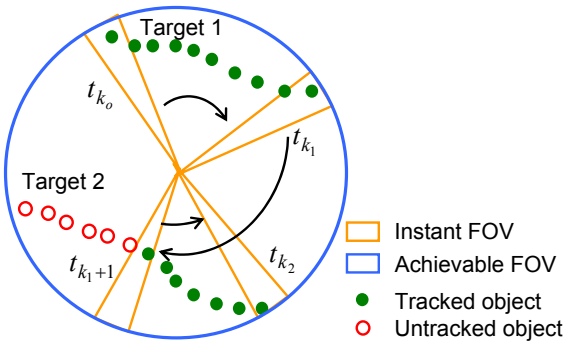


Figure 3.14. Illustration of the PTZ camera's instant and achievable FOVs. The discrepancy can be solved using an $\mathcal{M}/\mathcal{M}/1/1$ queuing system. Target 1 is tracked by the PTZ camera from t_{k_o} to t_{k_1}. As the maximum number of tracked objects $N_{obj} = 1$ has been achieved, target 2 cannot be processed immediately after it enters the PTZ camera's achievable FOV. Only after target 1 leaves the camera's achievable FOV, the PTZ camera can be directed to target 2 for object tracking.

$N_{obj} = 1$ has been achieved, target 2 cannot be processed until target 1 leaves the camera's achievable FOV. In Figure 3.14, target 2 is tracked from t_{k_1+1} to t_{k_2}. Figure 3.14 illustrates the aforementioned process, which is an exact $\mathcal{M}/\mathcal{M}/1/1$ queuing system. Therefore, in sensor planning, the achievable FOV with $N_{obj} = 1$ is sufficient to model PTZ cameras. In addition, the analysis of PTZ cameras is incorporated into a unified framework along with the static perspective and omnidirectional cameras. The only difference is the assumption regarding the maximum number of targets that can be tracked simultaneously. The maximum numbers of targets for a static camera and a PTZ camera are $N_{obj} \geq 1$ and $N_{obj} = 1$, respectively.

3.2.3 Additional constraints from performance requirements

Frequently special performance requirements are given. To meet these requirements, additional constraints need to be added. The coverage and resolution considerations correspond to priority areas that need complete coverage and/or with specified resolution. The frontal view requirement results from path constraints where there exist predefined paths within which the objects' movements are restricted.

There exist two approaches: direct constraint and adaptive weight, to impose these additional requirements. Considering the coverage requirement for example, the direct constraint approach finds the solution by imposing an extra constraint $A_1\mathbf{x} \geq \mathbf{b}_{C,o}$ where $\mathbf{b}_{C,o}$ represents the required coverage with $b_{C,o,i} = 1$ if the corresponding grid is to be covered and $b_{C,o,i} = 0$ otherwise. The adaptive weight approach assigns different weights $w_{1,i}$ to the grid points according to the coverage requirements. Larger weights are used if the corresponding grids need to be covered. The objective function then becomes:

$$c_i = w_{1,i}(c'_{1,i} > 0) + w_2(c'_{2,i} = 2) - w_3(c'_{3,i} > 1). \tag{3.31}$$

To incorporate the resolution requirements, we construct a matrix A_4 with $a_{4,ij} = 1$ if $M_{R,ij} \geq M_{R,o,i}$ and $a_{4,ij} = 0$ otherwise, where $M_{R,o,i}$ is the corresponding resolution requirement at the i^{th} grid point. The direct constraint approach is carried out by introducing an extra constraint $A_4\mathbf{x} \geq \mathbf{b}_{R,o}$ where $\mathbf{b}_{R,o}$ represents the required resolution with $b_{R,o,i} = 1$ if the corresponding grid needs the minimum resolution and $b_{R,o,i} = 0$ otherwise. In the adaptive weight approach the objective function becomes:

$$c_i = w_1(c'_{1,i} > 0) + w_2(c'_{2,i} = 2) - w_3(c'_{3,i} > 1) + w_{4,i}(c'_{4,i} > 0), \tag{3.32}$$

where $\mathbf{c}'_4 = A_4\mathbf{x}$ and $w_{4,i}$ are different weights allocated according to the resolution requirement.

In surveillance systems, a predefined path is commonly encountered. It is also preferred that a frontal view can be achieved sometime while pedestrians are moving

along this path. An example is the entrance areas where a frontal view of the pedestrian is desired when he or she enters the gate. We use the tangential direction of the middle line of the path as the average direction of the pedestrian's motion, as shown in Figure 3.15. Let the k^{th} point on the middle line be $P_{0,k}$ and its tangential direction be $\mathbf{u}_{P,k}$. The frontal view measure observed by the j^{th} camera at point $P_{i',k}$ along the line perpendicular to $\mathbf{u}_{P,k}$ is given by:

$$FV_{i'j} = \frac{(T_{C,j} - P_{i',k})^T \mathbf{u}_{P,k}}{\| T_{C,j} - P_{i',k} \| \| \mathbf{u}_{P,k} \|}. \tag{3.33}$$

Based on $FV_{i'j}$, we define a matrix A_5 with $a_{5,i'j} = 1$ if $FV_{i'j} \geq 0$ and $a_{5,i'j} = 0$ otherwise. Let $a_{5,ij} = 0$ for grid points outside the path. Finally the path constraint is incorporated into sensor planning by:

$$c_i = w_1(c'_{1,i} > 0) + w_2(c'_{2,i} = 2) - w_3(c'_{3,i} > 1) + w_{5,i}(c'_{5,i} > 0), \tag{3.34}$$

where $\mathbf{c}'_5 = A_5\mathbf{x}$ and $w_{5,i}$ are different weights allocated according to the frontal view requirement.

Note that although the coverage, resolution, and frontal view constraints are addressed separately, it is straightforward to combine any two terms or all three. The only modification is to add the corresponding terms. The adaptive weight approach is especially attractive because of its concise expression and speed of convergence.

3.3 Experimental results

In this section, we first validate the newly developed observation measure for

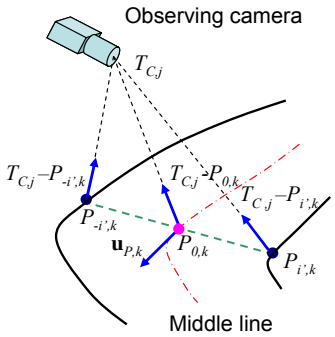


Figure 3.15. Illustration of how to compute the frontal view component with path constraints.

different types of cameras and then introduce our experimental methodology. Our experimental results using three floor plans are presented and compared with a reference algorithm proposed by Erdem and Sclaroff [Erdem06]. Three criteria are used to evaluate and compare the performances of various algorithms: coverage (C), handoff success rate (HSR), and frontal view percentage (FVP). For clear presentation, the reference algorithm is denoted as T1C and T2C for the Max-Coverage (Type 1) and Min-Cost (Type 2) problems, where C stands for coverage. Our sensor planning methods discussed in section 3.2.1 are denoted as T1H and T2H, where H stands for handoff. When the frontal view or path constraint is included, we refer to our methods described in section 3.2.3 as T1P and T2P, where P stands for the path constraint. Comparing the T1C (T2C) method with the T1H (T2H) method, an improved handoff success rate is expected. The major difference between the T1H and T1P (T2H and T2P) methods lies in that the path constraint is added in the T1P (T2P) method. Therefore an improved frontal view percentage is expected from the T1P and T2P methods. The algorithms for multiple dynamic targets discussed in section 3.2.2 are denoted as T1DO/T2DO, where D and O stand for dynamic occlusion and camera overload, respectively. Their performances are compared with the T1H method. Unlike the T1H method that suffers from a decreased handoff success rate as the number of targets in the environment increases, a maintained handoff success rate is expected from the T1DO method.

3.3.1 Experiments on observation measure

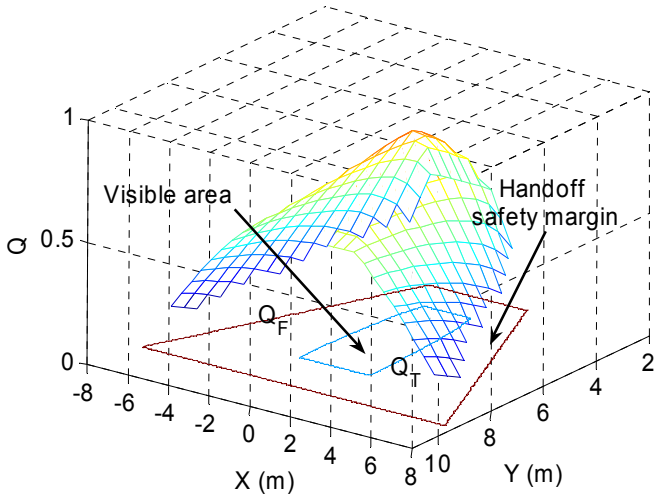
We begin this section with the discussion regarding the selection of parameters used in the definition of the observation measure. There are two sets of parameters: the normalization coefficients (α_R, α_D, and α_{FV}) and the importance weights (w_R, w_D, and w_{FV}). The goal of choosing the appropriate normalization coefficients is to provide a uniform comparison basis for different types of cameras and cameras with various intrinsic and extrinsic parameters. In so doing, sensor planning and camera handoff can be conducted disregarding the actual types of cameras involved. In general, we normalize the M_R and M_D components in the range of zero to one and the M_{FV} component in the range of minus one to one. For static perspective cameras, the maximum of M_R is achieved at $\hat{z}' = -T_Z / \tan\theta_T$. We have $M_{R,\max} = \alpha_R / \hat{z}'|_{\hat{z}'=-T_Z/\tan\theta_T} = 1$ and thus $\alpha_R = -T_Z / \tan\theta_T$. To normalize the M_D and M_{FV} components, we need $\alpha_D = 0.5$ and $\alpha_{FV} = 1$, respectively. As for omnidirectional cameras, the maximum of the M_R component is obtained by letting $\theta = 0 : M_{R,\max} = \frac{\alpha_R f Z}{Z^2 + R^2} \sum_{k=1,odd} \lambda_{\theta,k} k \theta^{k-1} \Big|_{\theta=0,R=0} = \frac{\alpha_R f \lambda_{\theta,1}}{Z}$. In consequence, we arrive at $\alpha_R = \frac{Z}{f\lambda_{\theta,1}}$. In the similar fashion, we set $\alpha_D = 1$ and $\alpha_{FV} = 1$ to normalize the M_D and M_{FV} components for omnidirectional cameras. Different from the selection of the normalization coefficients, which depend of the characteristics of the cameras used, the selection of the importance weights is purely application dependent.

In the following experiment, a static perspective camera is placed at T_C=[0 0 3m]^T looking down toward the ground plane at a tilt angle of -30°. Its pan angle is set to zero. The image size is 640×480. The camera's focal length is 21.0mm. Points are uniformly sampled in the ground plane (Z=0) with X in the range of -8m to 8m and Y in the range of 2m to 10m. Based on these parameters, the normalization coefficient of the M_R component is $\alpha_R = -3/\tan(-30^\circ) = 5.2$. As we mentioned before, a smaller decreasing/increasing rate of the M_D component is desired when the target is at long distance. In our implementation, we choose $\begin{cases} \beta_1 \mid T_Z / \tan\theta_T \mid + \beta_o = 1 \\ 2\beta_1 \mid T_Z / \tan\theta_T \mid + \beta_o = 0.5 \end{cases}$ and obtain $\beta_1 = -0.1$, $\beta_0 = 1.5$. In summary, the parameters used are listed as follows: $\alpha_R = 5.2$, $\alpha_D = 0.5$, $\beta_1 = -0.1$, $\beta_0 = 1.5$, $\alpha_{FV} = 1$, $w_R = 0.25$, $w_D = 0.75$, $w_{FV} = 0$. Figure 3.16(a) shows the observation measures for the perspective camera. The best observation area with the maximum observation measure is in the proximity of [0 5m 0]^T. As the object moves away from this area, the observation measure decreases. A higher penalty is given to the motion along the X-axis, the direction orthogonal to the camera's optical axis. The proposed observation measure gives a quantified evaluation of the tracking and observation suitability, which also agrees with our intuition and visual inspection.

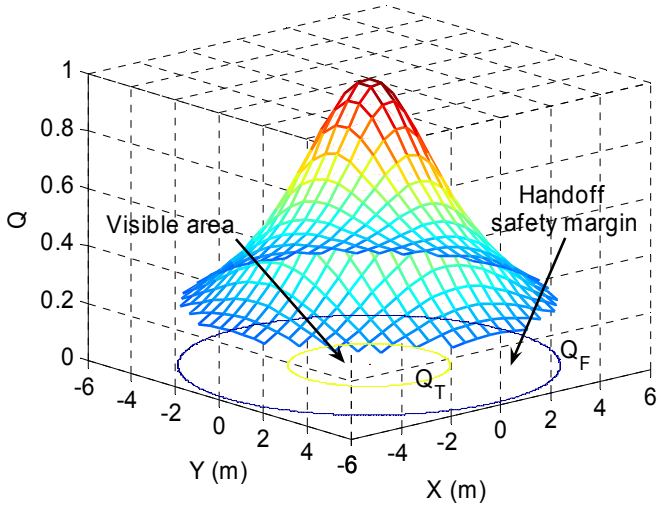
In the second simulation, an omnidirectional camera that follows the equidistance projection model is placed at T_C=[0 0 3m]^T overlooking an area with (X, Y) in the range of -6m to 6m. The image size is 640×640. The normalization coefficient for the resolution component is given by $\alpha_R = \frac{6}{640} = 9.4\times10^{-3}$. Other parameters used are listed as follows: $\alpha_D = 1$, $\alpha_{FV} = 1$, $w_R = 0.25$, $w_D = 0.75$, $w_{FV} = 0$. The resulting observation measure is shown in Figure 3.16(b). A radial symmetric shape is obtained, which coincides with the characteristics of an omnidirectional camera.

3.3.2 Experimental methodology

The floor plans used in this section are shown in Figure 3.17. The floor plan in Figure 3.17(a) represents two types of environments commonly encountered in practical surveillance: space with obstacles (region A illustrated in yellow) and open space where pedestrian can move freely (region B illustrated in green). Region B is deliberately included because it imposes more challenges on camera placement when considering handoff success rate. Camera handoff is relatively easier when there is a predefined path compared with the scenarios where subjects can move freely, since camera handoff may be triggered at any point in the camera's FOV. Figure 3.17(b) illustrates an environment with a predefined path where workers proceed in a predefined sequence. The floor plan of an outdoor parking lot is also included to evaluate the performance of the proposed algorithms for large scale environments. The dimensions of the parking lot are about 50m×100m. In the following experiments, we refer to theses plans as plan A, B, and C. In our experiments, static perspective cameras are placed along the walls of the environment while omnidirectional and PTZ cameras can be mounted on the ceiling and at sampled grid points with an interval of 1m.

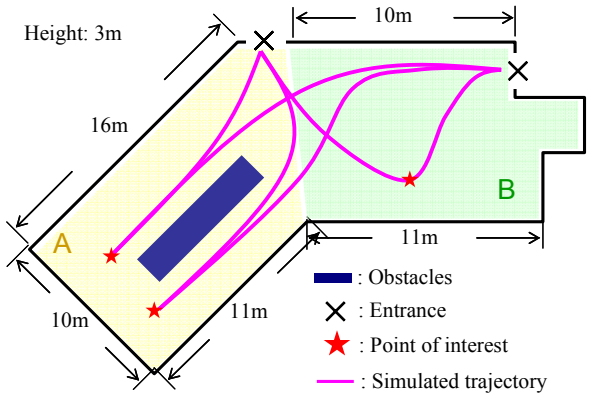


(a)

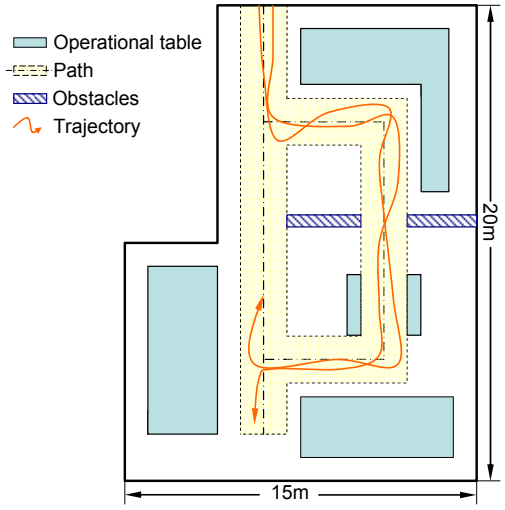


(b)

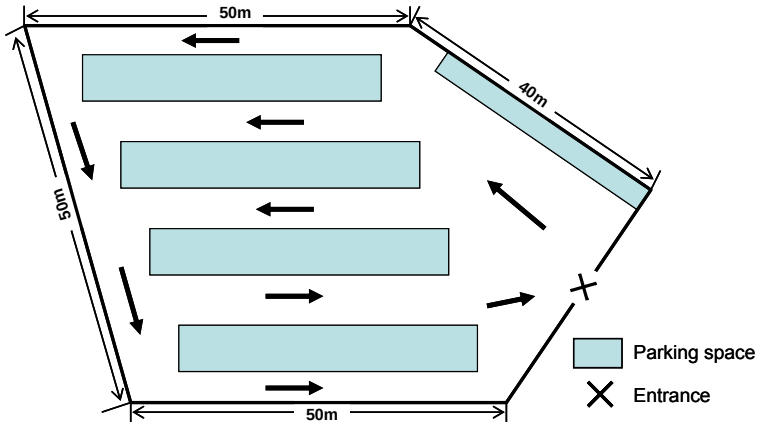
Figure 3.16. Graphical illustration of the observation measure and handoff safety margin for (a) perspective and (b) omnidirectional cameras.



(a)



(b)



(c)

Figure 3.17. Tested floor plans. Two office floor plans: (a) without path constraints and (b) with path constraints. (c) A floor plan of an outdoor parking lot.

To obtain a statistically valid estimation of the handoff success rate, simulations are carried out to enable a large amount of tests under various conditions. A pedestrian behavior simulator [Antonini06, Pettre02] is implemented so that we could achieve a close resemblance to experiments in real environments and in turn an accurate estimation of the handoff success rate. Interested readers can refer to the original papers for details. In our experiments, the arrival of the pedestrian follows a Poisson distribution with an average arrival rate of 0.01 persons per second. The average walking speed is 0.5m per second. Several points of interest are generated randomly to form a pedestrian trace. Figure 3.17 also depicts some randomly generated pedestrian traces. Handoff success rate and frontal view percentage are obtained from simulation results of 300 randomly generated traces.

3.3.3 Experiments on sensor planning

In the following experiments, w_1, w_2, and w_3 are set to 1, 2, and 5. The failure and trigger thresholds are 0 and 0.6, respectively. Since for both indoor floor plans the required visible distance is about 10m and the height is 3m, the same pair of tilt angle and focal length can be used for static perspective cameras with $f = 21.0mm$ and $\theta_T = -30^\circ$.

Figure 3.18 illustrates the experimental results for floor plan A using static perspective cameras to solve the Max-Coverage problem. Our T1H approach chooses a camera positioning scheme with a slightly decreased coverage from 81.6% to 74.7%. However, the HSR is improved substantially from 23.2% to 87.4%. An example trace is also shown in Figures 3.18(c) and (d). As expected, if only coverage is considered, insufficient overlapped FOVs are kept between adjacent cameras, leading to two handoff failures as observed in Figure 3.18(c). In comparison, given the camera placement optimized by the T1H method, the target is tracked continuously with three successful handoffs as shown in Figure 3.18(d).

As expected, a considerably improved HSR is also achieved for floor plan B as shown in Figure 3.19. In addition, we add the frontal view criterion with w_5=5 and test the T1P method. The FVP is elevated from 28.7% to 93.5%. From Figures 3.19(b) and (c), we could see that the cameras are oriented toward the direction of the predefined path after introducing the frontal view constraint.

The Min-Cost problem imposes additional requirements on the overall coverage, which leaves less freedom in the optimization process to achieve the maximum HSR. As Figure 3.20 demonstrates, the overall coverage is constrained to be above 80%, which results in a decrease in HSR from 87.4% to 68.5%. However, with similar coverage (T2H: 81.5% vs. T1C: 81.6%), our T2H algorithm is able to achieve a much higher HSR (68.5%) than the conventional T1C approach (23.2%). Figure 3.21 demonstrates similar performance comparison of the T2H and T2P methods for floor plan B.

Figures 3.22 and 3.23 demonstrate the experimental results for PTZ cameras. The HSR is elevated from 48.7% to 100% at the cost of a marginal decrease in coverage from 100.0% to 99.5% when comparing the performance of the T1C and T1H methods for floor plan B. Our placement algorithm works properly for floor plan C, an example of large scale environments. The T1H method generates an HSR of 99.9% and a coverage

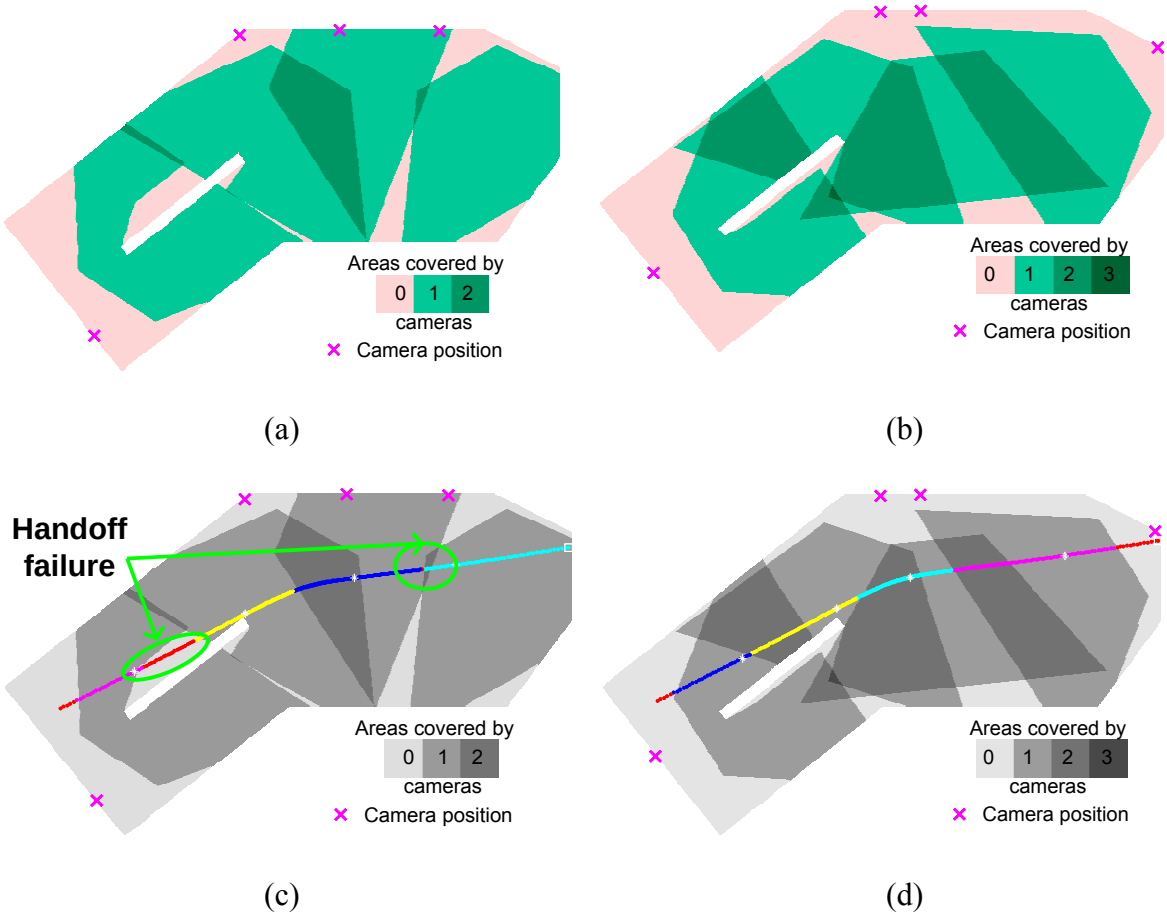
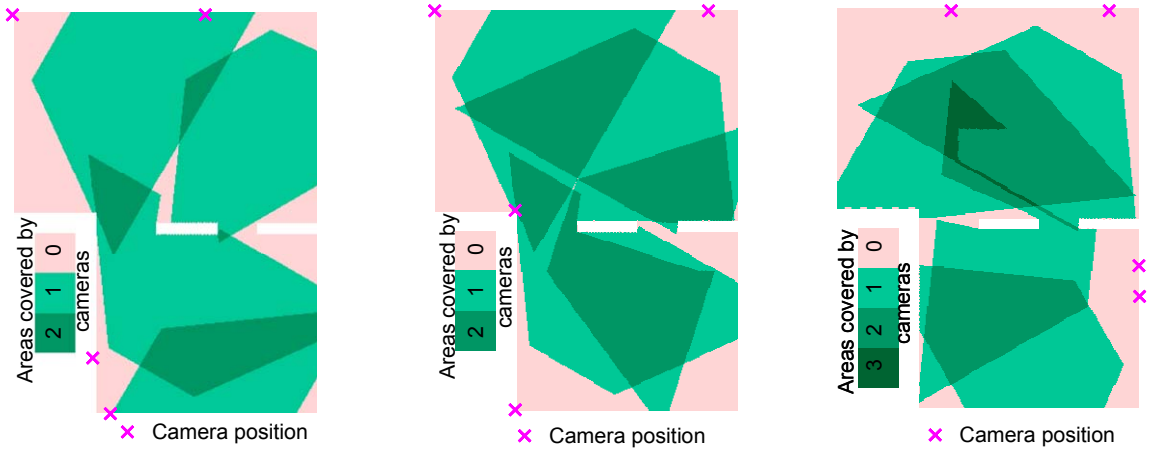
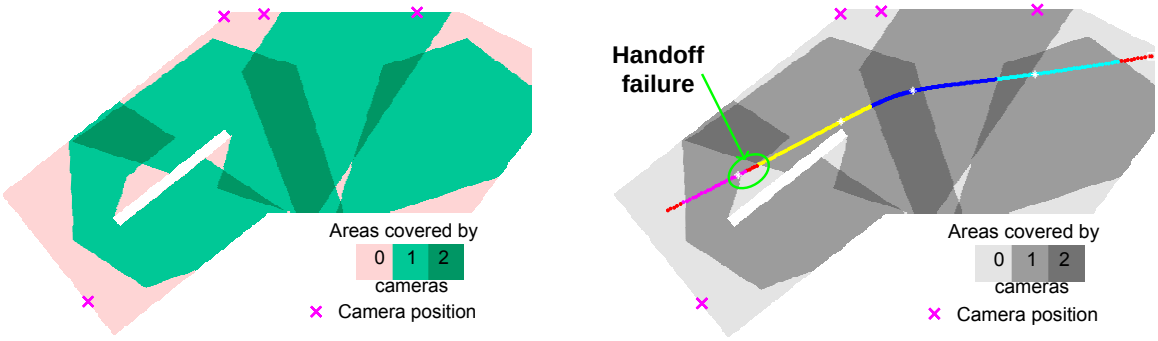


Figure 3.18. Optimal camera positioning of floor plan A for the Max-Coverage problem using perspective cameras (a) T1C (C: 81.6 %, HSR: 23.2%) and (b) T1H (C: 74.7%, HSR: 87.4%). An example trace: two handoff failures in (c) and three successful handoffs in (d).



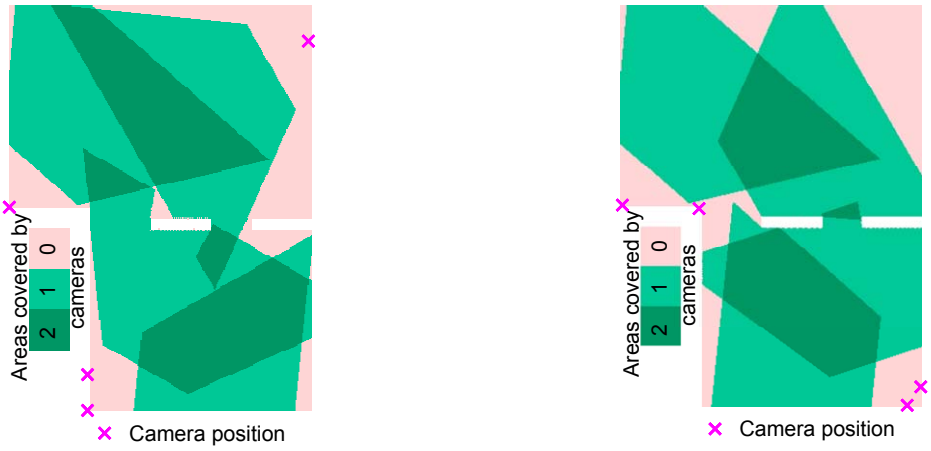
(a) (b) (c)

Figure 3.19. Optimal camera positioning of floor plan B for the Max-Coverage problem using perspective cameras (a) T1C (C: 84.8%, HSR: 6.0%, FVP: 67.7 %), (b) T1H (C: 74.7%, HSR: 56.9 %, FVP: 28.7%), and (c) T1P (C: 72.1%, HSR: 58.0%, FVP: 93.5%).



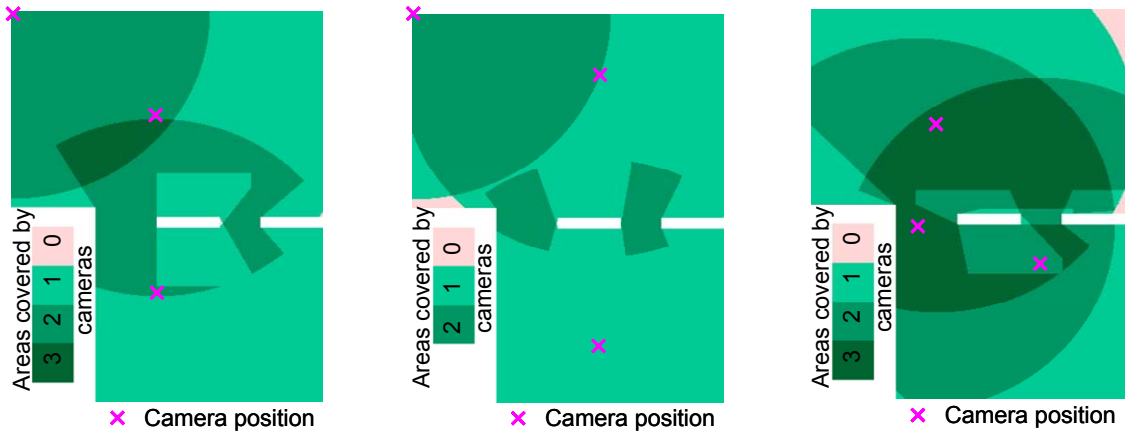
(a) (b)

Figure 3.20. Optimal camera positioning of floor plan A for the Min-Cost problem using perspective cameras (C≥80%). (a) T2H (C: 81.5%, HSR: 68.5%). (b) An example trace.



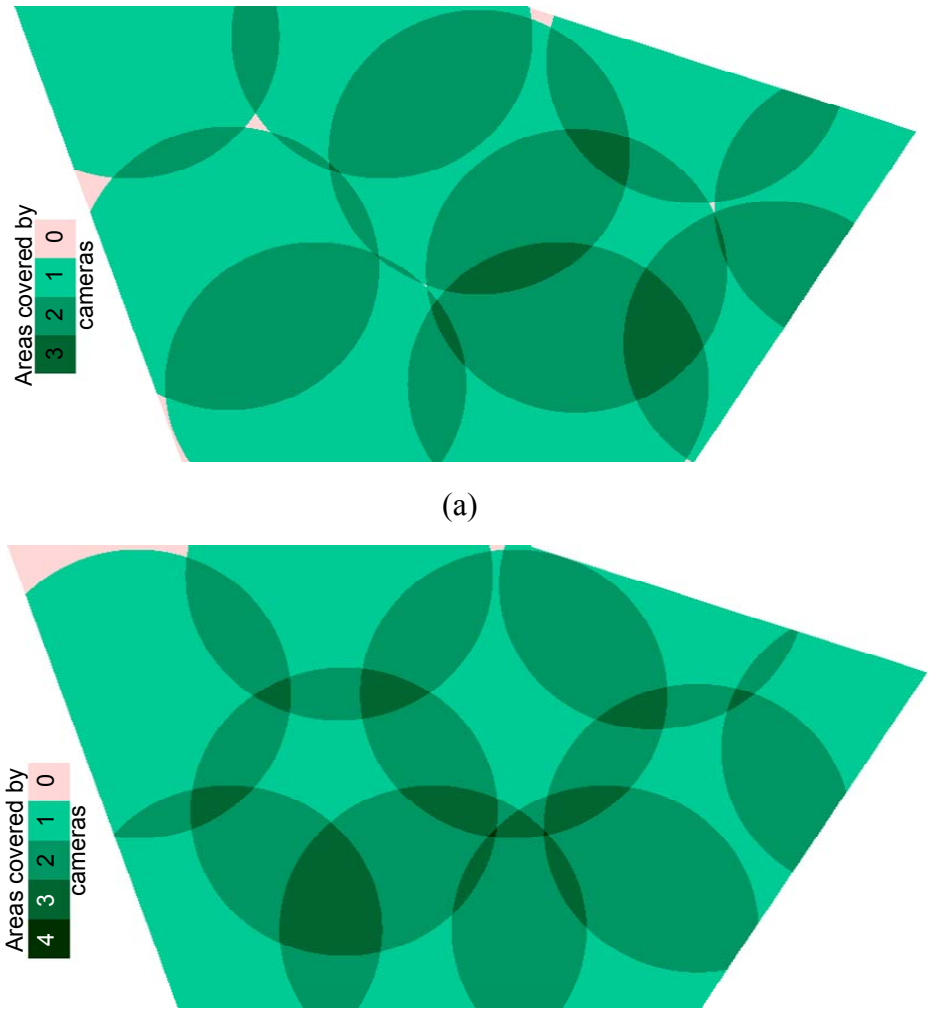
(a) (b)

Figure 3.21. Optimal camera positioning of floor plan B for the Min-Cost problem using perspective cameras (C≥80%): (a) T2H (C: 81.3%, HSR: 43.7%, FVP: 41.0%) and (b) T2P (C: 81.6 %, HSR: 47.1 %, FVP: 69.0%).



(a) (b) (c)

Figure 3.22. Optimal camera positioning of floor plan B for the Max-Coverage problem using PTZ cameras: (a) T1C (C: 100.0%, HSR: 48.7%, FVP: 52.5%), (b) T1H (C: 99.5%, HSR: 100.0%, FVP: 53.4%), and (c) T1P (C: 99.0%, HSR: 100.0%, FVP: 71.1%).



(a)

(b)

Figure 3.23. Optimal camera positioning of floor plan C for the Max-Coverage problem using PTZ cameras: (a) T1C (C: 99.5%, HSR: 73.5%) and (b) T1H (C: 99.2%, HSR: 99.9%).

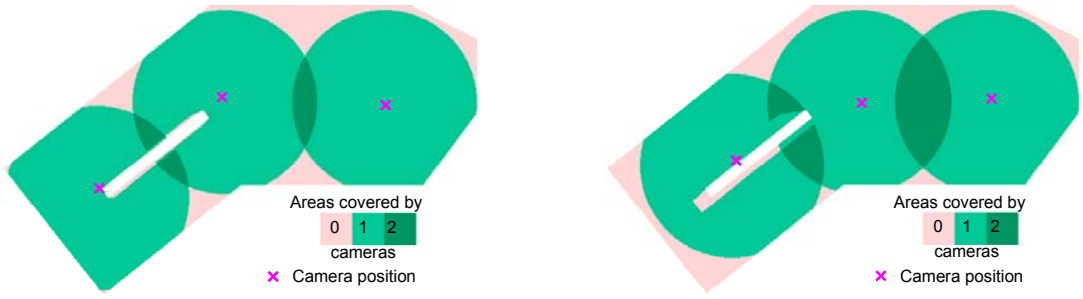
of 99.2% in comparison with the HSR of 73.5% and coverage of 99.5% from the T1C method.

In parallel, experiments are conducted using omnidirectional cameras. To cover a radius of 6m at a height of 3m, the chosen focal length is 15.4mm. Figures 3.24 and 3.25 show the optimal camera placement. At the cost of 2.4% decrease in coverage, the HSR increases from 52.8% to 79.0% for floor plan A. Different from perspective cameras which can look into a particular direction for a frontal view of the target, omnidirectional cameras have a 360°×90° view. Therefore, the improvement in FVP from imposing the frontal view constraint is not substantial, indicated by an increase of 4.2% from T1H to T1P.

Table 3.1 summarizes the performance comparison between the proposed algorithms and the reference algorithm described by Erdem and Sclaroff [Erdem06]. Consistent observations are obtained from experiments using three floor plans and three types of cameras. Compared with the reference algorithm, our algorithms produce considerably improved HSR and FVP at the cost of slightly decreased coverage. This amount of decrease in coverage is inevitable in order to maintain overlapped FOVs between adjacent cameras required by continuous and automated tracking given a fixed number of cameras. The ratio between the increase in HSR and the decrease in coverage $\Delta HSR/|\Delta C|$ describes the advantage of our algorithms. For the Max-Coverage problem, every 1% decrease in coverage results in a 4% to 10% increase in HSR. An even higher improvement rate can be achieved for the Min-Cost problem. The efficiency of the proposed algorithms in balancing the overall coverage and sufficient overlapped FOVs becomes evident. Furthermore, our algorithms can handle additional constraints as well, such as the frontal view requirement. The resulting T1P and T2P algorithms are able to maintain a similar improvement rate in HSR as the T1H method with further improved FVP.

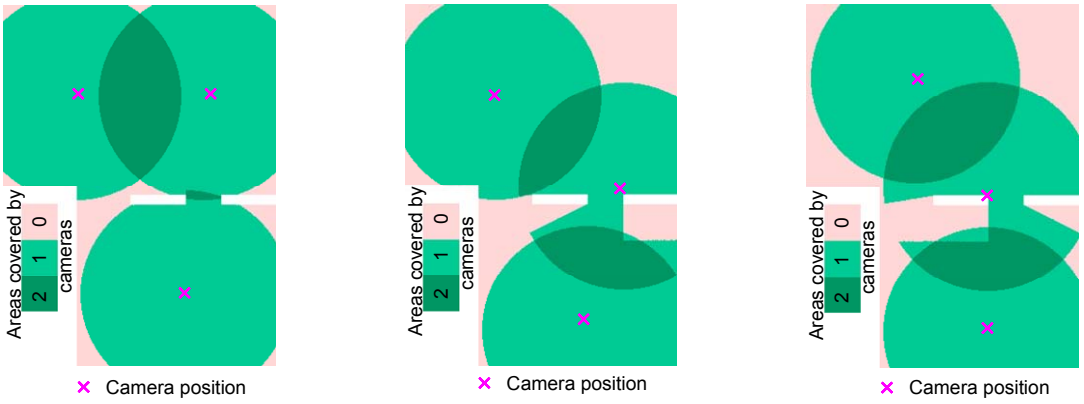
The conventional sensor planning methods achieve a camera placement with a maximized coverage. In such a system, although it can be seen, the target cannot be consistently labeled or recognized as the same identity across different cameras because of handoff failures resulting from insufficient overlapped FOVs. The resulting camera placement cannot support automated and persistent surveillance since the tracked or identified target trajectories are disjoint at the junction areas of adjacent cameras. In contrast, our sensor placement ensures a continuous and consistently labeled trajectory. The slightly decreased coverage can be easily compensated for by adding an additional camera. The cost of an extra camera is acceptable in comparison with a system with inherent disability of persistent and continuous tracking.

Finally, we study the performance of our sensor planning algorithms for environments with multiple dynamic targets. PTZ cameras are used to include both problems of dynamic occlusion and camera overload. The corresponding importance weights are w_4=5 and w_5=5. Different target densities are tested to study their influence on camera placement. Figure 3.26 (a) shows the camera placement obtained from the T1H method for floor plan B. Figures 3.26(b) and (c) depict the camera placement with different target densities, K_o. A larger K_o suggests an environment with a higher target density and leads to a camera placement with more overlapped FOVs between adjacent cameras so that the tracked target has more freedom to be transferred to another camera

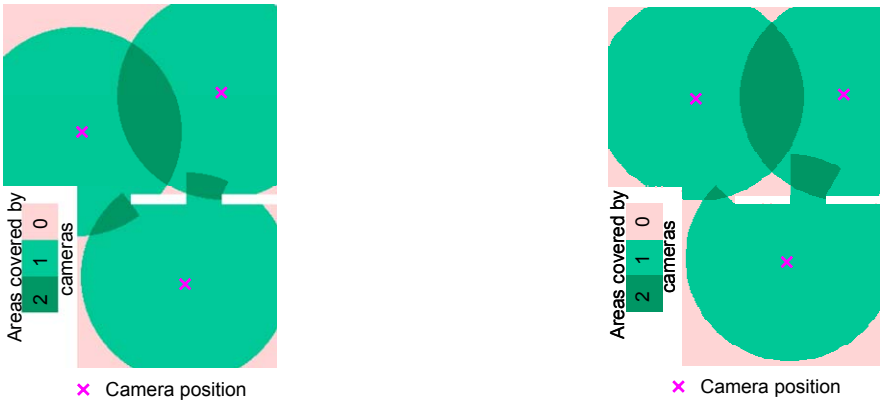


(a) (b)

Figure 3.24. Optimal camera positioning of floor plan A for the Max-Coverage problem using omnidirectional cameras (a) T1C (C: 88.4 %, HSR: 52.8%) and (b) T1H (C: 86.0%, HSR: 79.0%).



(a) (b) (c)

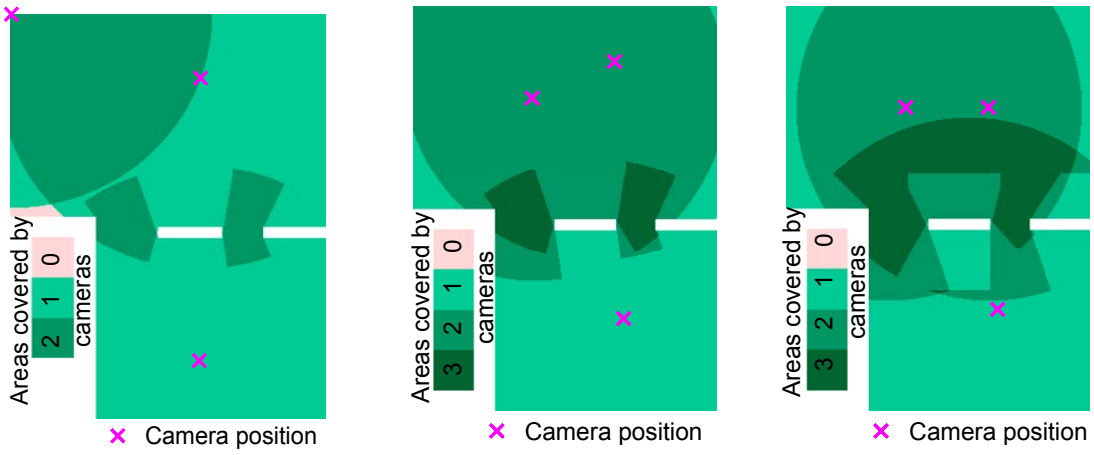


(d) (e)

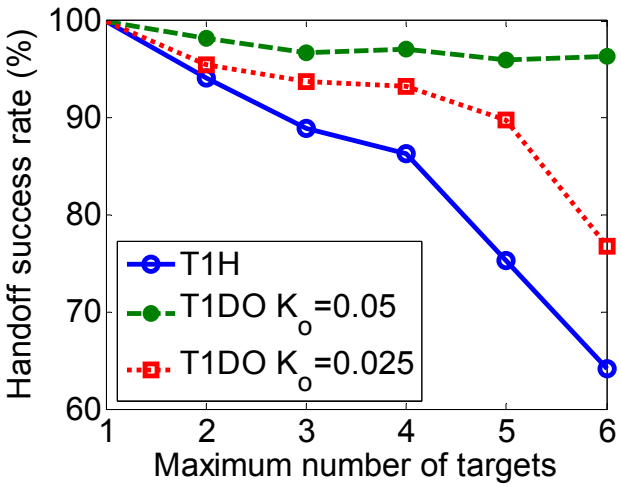
Figure 3.25. Optimal camera positioning of floor plan B using omnidirectional cameras. The Max-Coverage problem: (a) T1C (C: 92.1%, HSR: 50.0%, FVP: 49.9%), (b) T1H (C: 81.5 %, HSR: 92.6%, FVP: 53.4%), and (c) T1P (C: 80.0%, HSR: 100.0%, FVP: 57.6%). The Min-Cost problem (C≥90%): (d) T2H (C: 91.2%, HSR: 52.2%, FVP: 45.7%) and (e) T2P (C: 90.7%, HSR: 100.0%, FVP: 53.4%).

Table 3.1. System performance comparison.

Floor plan A (30m×10m)					
Camera	Method	C	HSR	ΔHSR/|ΔC|	
Perspective	T1C/T2C	81.6	23.2		
	T1H	74.7	87.4	9.3	
	T2H (C>80%)	81.5	68.5	453	
Omnidirectional	T1C/T2C	88.4	52.8		
	T1H/T2H	86.0	79.0	10.9	
Floor plan B (20m×15m)					
	Method	C	HSR	ΔHSR/|ΔC|	FVP
Perspective	T1C/T2C	84.8	6.0		67.7
	T1H	74.7	56.9	5.0	28.7
	T1P	72.1	58.0	4.1	93.5
	T2H (C>80%)	81.3	43.7	10.8	41.0
	T2P	81.6	47.1	12.8	69.0
Omnidirectional	T1C/T2C	92.1	50.0		49.9
	T1H	81.5	92.6	4.0	53.4
	T1P	80.0	100.0	4.1	57.6
	T2H (C>90%)	91.2	52.2	2.4	45.7
	T2P	90.7	100.0	35.7	53.4
PTZ	T1C/T2C	100.0	48.7		52.5
	T1H	99.5	100.0	102.6	53.4
	T1P	99.0	100.0	51.3	71.1
Floor plan C (50m×100m)					
Camera	Method	C	HSR	ΔHSR/|ΔC|	
PTZ	T1C/T2C	99.5	73.5		
	T1H	99.2	99.9	88.0	



(a) (b) (c)



(d)

Figure 3.26. Sensor planning results considering the problems of dynamic occlusion and camera overload. The optimal camera positioning of floor plan B for the Max-Coverage problem using PTZ cameras: (a) T1H, (b) T1DO with K_o=0.025, and (c) T1DO with K_o=0.05. (d) System performance comparison based on handoff success rate with various target densities. The target density is described by the maximum number of targets to be tracked simultaneously in the environment.

when experiencing dynamic occlusion and/or camera overload. The advantage of the T1DO method over the T1H method becomes more conspicuous when we look into the handoff success rate with respect to various target densities, as shown in Figure 3.26(d). Note that in this plot the maximum number of targets is used to describe the target density instead of K_o for a more intuitive view of the target's distribution in the environment. For the T1H method, the HSR drops gradually from 100% to 64.7% as the maximum number of targets increases from one to six. On the contrary, for the T1DO method, the HSR is maintained within 90% for camera placement with K_o=0.025 (K_o=0.05) till the maximum number of targets reaches four (six). As expected, the camera placement with a higher target density K_o=0.05 yields a more robust performance in more clustered environments.

4 Size preserving tracking

To achieve size preserving tracking, in addition to controlling the camera's pan and tilt motions to keep the object of interest in the camera's FOV, the camera's focal length is adjusted automatically to compensate for the changes in the target's image size caused by the relative motion between the camera and the target. The estimation accuracy of these changes determines the effectiveness of the resulting zoom control. Considering accuracy, computational complexity, and robustness to image noise and based on the survey presented in section 2.2, the target depth based algorithm is selected. The existing method of choice applies structure from motion (SFM) based on the weak perspective projection model. We propose a target scale estimation algorithm with a linear solution based on the more advanced paraperspective projection model, which improves the accuracy of scale estimation by considering center offset. Another key issue in the target depth based algorithms is the separation of foreground and background features, especially when composite camera (pan/tilt/zoom) and target motions are involved. We also design a fast foreground/background segmentation algorithm, the affine shape method. The resulting segmentation automatically adapts to the target's 3D geometry and motion and is able to accommodate a large amount of off-plane rotation, which most existing segmentation algorithms find difficult to achieve.

The remainder of this chapter is organized as follows. Section 4.1 gives a brief overview of our size preserving tracking algorithm. The proposed scale estimation and foreground segmentation algorithms along with experimental results are presented in sections 4.2 and 4.3, respectively.

4.1 Algorithm description

Figure 4.1 shows the algorithm's flow chart including feature detection and matching, foreground and background segmentation, SFM, gaze point estimation, and target scale estimation. In our implementation, features (image corners) are detected and tracked based on Shi's method [Shi94] and the pyramidal Lucas-Kanade tracker [Bouguet00], respectively. Conventional factorization approach is used for recovering structure and motion [Tomasi92]. In this chapter, we will focus on two decisive steps: target scale estimation and foreground/background segmentation.

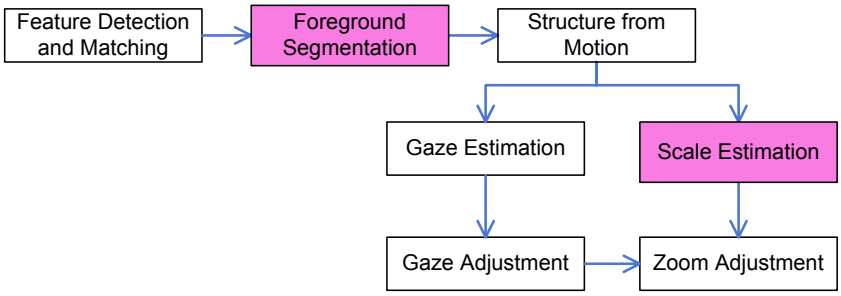


Figure 4.1. Flow chart of the size preserving tracking algorithm.

Size preserving tracking consists of two fundamental functionalities, target pursuing and target scale preservation. The target pursuing unit controls the camera's pan and tilt angles, so that the target remains in the camera's FOV while the target scale preservation unit provides the guidelines to vary the camera's focal length. To maintain a constant target image size, we need to estimate and predict the changes in target scale based on preceding frames and issue the proper zoom commands to counteract these changes. The scale estimation algorithm described in [Tordoff04] utilizes the weak perspective projection model, which is a highly abstracted simplification of the real imaging process. To improve the accuracy in target scale estimation, we examine the possibility of using more advanced projection models.

Meanwhile, we are interested in fast and efficient foreground and background segmentation schemes for real-time applications. In this chapter, we describe a segmentation algorithm based on affine shapes. The prominent advantage of the proposed segmentation algorithm is its fast implementation. In addition, at the cost of linear computations, the resulting algorithm is able to achieve satisfactory accuracy and robustness to image noise and off-plane rotation.

4.2 Scale estimation

In literature, the perspective and affine projection models are two major types of projections used to describe the imaging process. In the perspective projection model, an image point $\mathbf{p}$ is projected from a scene point P by: $\begin{bmatrix}\mathbf{p}\\ z\end{bmatrix} = M_P \begin{bmatrix}P\\ 1\end{bmatrix} = K\begin{bmatrix}R & \mathbf{t}\end{bmatrix}\begin{bmatrix}P\\ 1\end{bmatrix}$, where z denotes the depth, M_P, a 3×4 matrix, is the perspective projection matrix, and K, R, $\mathbf{t}$ are the camera intrinsic matrix, rotation matrix, and translation vector, respectively. If the target's relief is small enough compared to its distance from the observing camera, affine projection models can be used to approximate the imaging process. Affine projection is characterized by the following equation: $\mathbf{p} = M_A \begin{bmatrix}P\\ 1\end{bmatrix}$, where M_A, a 2×4 matrix, is the affine

projection matrix. The weak perspective and paraperspective projection models are two examples of the affine projection model.

The weak perspective projection of a point P is constructed in two steps. (1) The scene point P is first projected orthographically onto a point P' on the reference plane Π_r, which is parallel to the image plane Π and passes the target's center of mass. (2) A pin-hole model, which corresponds to a scaling of the coordinates, projects P' onto a point $\mathbf{p}$ in Π. If we denote the camera's intrinsic parameters as follows, skew: s, aspect ratio: α, and focal length: f, and the target's center of mass $[X_r\ Y_r\ Z_r]^T$, the projection matrix is given by:

$$M_A = \frac{f}{Z_r}\begin{bmatrix} 1 & s \\ 0 & \alpha \end{bmatrix}\begin{bmatrix} R_2 & \mathbf{t}_2 \end{bmatrix}, \tag{4.1}$$

where R_2 /$\mathbf{t}_2$ denotes the matrix/vector consisting of the first two rows of the matrix R /vector $\mathbf{t}$. In the weak perspective projection, the target's image simply translates when the target translates parallel to the image plane. However, under the perspective projection, the target's image presents a different view, which may introduce changed image size. This amount of change in the target's image size is determined by the center offset, target relief, and target depth.

The paraperspective projection evolves from the weak perspective projection. It takes into account both the distortions associated with the center offset and possible variations in target depth. It yields a closer approximation of the perspective projection by modeling the position effect. In the meanwhile, it also maintains some of the linear properties of the weak perspective projection, which makes it attractive to our intended applications. Similarly, the paraperspective projection involves two steps. The scene point P is first projected onto a point P' of Π_r along the direction of the line connecting the target's center of mass and the camera's optical center. The second step follows that of the weak perspective projection model. The projection matrix can be expressed as:

$$M_A = \frac{f}{Z_r}\left[\begin{bmatrix} 1 & s & x_r/f \\ 0 & \alpha & y_r/f \end{bmatrix} R \quad \begin{bmatrix} 1 & s \\ 0 & \alpha \end{bmatrix}\mathbf{t}_2\right]. \tag{4.2}$$

The image of the target's center of mass, $[x_r\ y_r]^T$, appears in the projection matrix. Letting x_r=y_r=0, we have the same expression as the weak perspective projection.

4.2.1 Theoretical derivation

For the purpose of size preserving tracking, our concern is the ratio between f and Z_r, $\rho_i = (f/Z_r)_i$, in the i^{th} frame. The camera's focal length is adjusted for a constant ρ_i and thus a constant target image size. The objective of scale estimation is to compute this ratio based on matched features in consecutive frames.

Assume that the target's feature points are tracked throughout the sequence. Under the affine projection, the unregistered j^{th} image point in the i^{th} frame, $\mathbf{p}_{ij,}$ is projected from

the scene point P_j by $\mathbf{p}_{ij} = M_i P_j + \mathbf{m}_i$. The registered points are formed by removing the estimated translation, $\mathbf{p}'_{ij} = \mathbf{p}_{ij} - \mathbf{m}_i$, where $\mathbf{m}_i = \frac{1}{J}\sum_{j=1}^{J} \mathbf{p}_{ij}$ and J is the total number of features. From J registered point correspondences established over I frames, Tomasi and Kanade [Tomasi92] recovered the affine structure and motion in a batch mode from the SVD of the $2I \times J$ registered measurement matrix:

$$W_i = \begin{bmatrix} \mathbf{p}'_{i-I+1,1} & \cdots & \mathbf{p}'_{i-I+1,J} \\ \vdots & \cdots & \vdots \\ \mathbf{p}'_{i,1} & \cdots & \mathbf{p}'_{i,J} \end{bmatrix} = \begin{bmatrix} \mu_1 & \cdots & \mu_{2I} \end{bmatrix} \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{2I} \end{bmatrix} \begin{bmatrix} \nu_1 & \cdots & \nu_{2I} \end{bmatrix}^T, \tag{4.3}$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{2I}$ in descending order and used its rank-3 property to find the optimal affine projection matrices M_i, M_{i-1}, …, M_{i-I+1} and structure $P_{i,j}$:

$$\begin{bmatrix} M_{i-I+1} \\ \vdots \\ M_i \end{bmatrix} = \begin{bmatrix} \sigma_1\mu_1 & \sigma_2\mu_2 & \sigma_3\mu_3 \end{bmatrix} \\ \begin{bmatrix} P_{i,1} & \cdots & P_{i,J} \end{bmatrix} = \begin{bmatrix} \nu_1 & \nu_2 & \nu_3 \end{bmatrix}^T \tag{4.4}$$

Let H be the non-singular matrix relating the recovered affine structure to the metric structure. The following relation holds: $M_{E,i'} = M_{i'} H$, $i' = i, i-1, \ldots, i-I+1$. Assuming zero skew and unit aspect ratio and exploring the orthogonality of the rotation matrix R, we have:

$$M_{E,i'} M_{E,i'}{}^T = (M_{i'} H)(M_{i'} H)^T = \left(\frac{f}{Z_r}\right)_{i'}^2 \begin{bmatrix} 1 + \frac{x_{r,i'}^2}{f^2} & \frac{x_{r,i} y_{r,i'}}{f^2} \\ \frac{x_{r,i'} y_{r,i'}}{f^2} & 1 + \frac{y_{r,i'}^2}{f^2} \end{bmatrix}. \tag{4.5}$$

Let

$$D_{i'} = \begin{bmatrix} \left(m_{11}^{i'}\right)^2 & 2m_{11}^{i'}m_{12}^{i'} & 2m_{11}^{i'}m_{13}^{i'} & \left(m_{12}^{i'}\right)^2 & 2m_{12}^{i'}m_{13}^{i'} & \left(m_{13}^{i'}\right)^2 \\ m_{11}^{i'}m_{21}^{i'} & m_{11}^{i'}m_{22}^{i'} + m_{12}^{i'}m_{21}^{i'} & m_{11}^{i'}m_{23}^{i'} + m_{13}^{i'}m_{21}^{i'} & m_{12}^{i'}m_{22}^{i'} & m_{12}^{i'}m_{23}^{i'} + m_{13}^{i'}m_{22}^{i'} & m_{13}^{i'}m_{23}^{i'} \\ \left(m_{21}^{i'}\right)^2 & 2m_{21}^{i'}m_{22}^{i'} & 2m_{21}^{i'}m_{23}^{i'} & \left(m_{22}^{i'}\right)^2 & 2m_{22}^{i'}m_{23}^{i'} & \left(m_{23}^{i'}\right)^2 \end{bmatrix}, \tag{4.6}$$

with $M_{i'} = \begin{bmatrix} m_{11}^{i'} & m_{12}^{i'} & m_{13}^{i'} \\ m_{21}^{i'} & m_{22}^{i'} & m_{23}^{i'} \end{bmatrix}$ and $HH^T = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_2 & h_4 & h_5 \\ h_3 & h_5 & h_6 \end{bmatrix}$. Assuming a unit scale in the $(i\text{-}I\text{+}1)^{th}$ frame and with known or estimated f, we arrive at the following linear equations to solve for $\mathbf{h} = \begin{pmatrix} h_1 & h_2 & h_3 & h_4 & h_5 & h_6 \end{pmatrix}$ and target scales, ρ_i, ρ_{i-1}, …, ρ_{i-I+2}:

$$D\begin{bmatrix}\mathbf{h} & \rho_{i-I+2}^2 & \cdots & \rho_i^2\end{bmatrix}^T = \begin{bmatrix}1+\frac{x_{r,i-I+1}^2}{f^2} & \frac{x_{r,i-I+1}y_{r,i-I+1}}{f^2} & 1+\frac{y_{r,i-I+1}^2}{f^2} & \mathbf{0}\end{bmatrix}^T, \tag{4.7}$$

where

$$D = \begin{bmatrix} D_{i-I+1} & 0 & 0 & \cdots & 0 \\ D_{i-I+2} & \mathbf{d}_{i-I+2} & 0 & \cdots & 0 \\ D_{i-I+3} & 0 & \mathbf{d}_{i-I+3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ D_i & 0 & 0 & \cdots & \mathbf{d}_i \end{bmatrix}, \tag{4.8}$$

with

$$\mathbf{d}_{i'} = \begin{bmatrix} -\left(1+\frac{x_{r,i'}^2}{f^2}\right) & -\frac{x_{r,i'}y_{r,i'}}{f^2} & -\left(1+\frac{y_{r,i'}^2}{f^2}\right)\end{bmatrix}^T. \tag{4.9}$$

The image of the target's center of mass, $x_{r,i'}$ and $y_{r,i'}$, can be obtained from the estimated translation $\mathbf{m}_{i'}$ [Poelman97]. The vector $\mathbf{h}$ has six unknowns and for I frames there are (I-1) $\rho_{i'}$. For each frame we can obtain three constrains and need $3I \geq 6 + I - 1$ or equivalently $I \geq 3$ frames to solve the above equations. The resulting algorithm requires similar computations as those based on the weak perspective projection model. The paraperspective projection model takes the target's position into consideration and can produce more accurate scale estimates when the target's image is drifted away from the image center.

We also designed scale estimation algorithms using the perspective projection model to relax the affine assumption that the target should be at a distance sufficiently large compared to its relief [Yao06C]. To achieve linear computations, the target's motion is restricted to a planar motion, representative of the motion of traffic and pedestrian. However, the use of the perspective projection model usually requires a final bundle adjustment to refine the reconstructed motion and structure. Even though the linear solution is mathematically valid, it is subject to image noise and the resulting estimation is unstable especially when composite camera and target motions are involved. In this chapter, we use the scale estimation algorithm based on the perspective projection model to study the distance constraint imposed by the affine assumption in controlled environments.

4.2.2 Experimental results

Offline and real-time pedestrian sequences are captured. The offline sequence is collected by a Sony camcorder DCR-TRV730 with a constant zoom and is used for evaluating the performance of the proposed scale estimation algorithms. The real-time

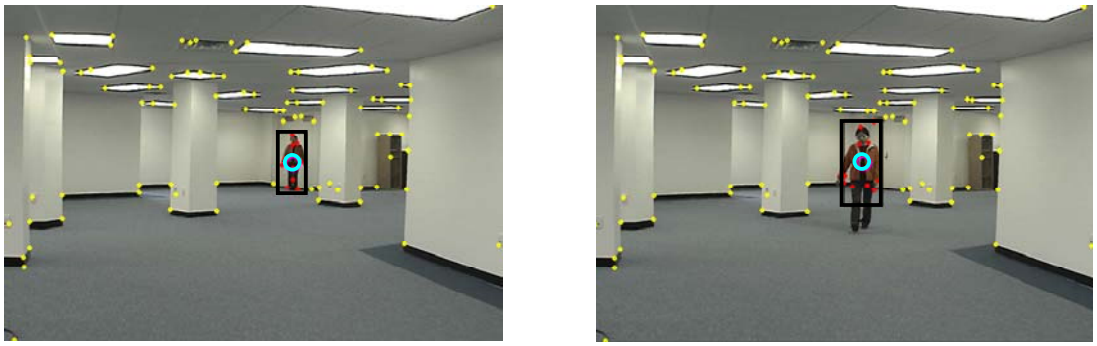
sequence is collected by a Pelco Spectra III SE series dome system. Real-time pan/tilt/zoom commands are issued to track the target and maintain a constant target image size. The scale estimation algorithms based on the weak perspective [Torrod04] and perspective [Yao06C] projection models are also implemented and their performance serves as a comparison reference.

Figure 4.2 shows sample frames and performance comparison from an offline pedestrian sequence. The target walks at a normal speed toward the camera from a distance of 15m to a distance of 5m. We manually measured the target's image size and used it as a reference to evaluate the performance of the algorithms based on various projection models.

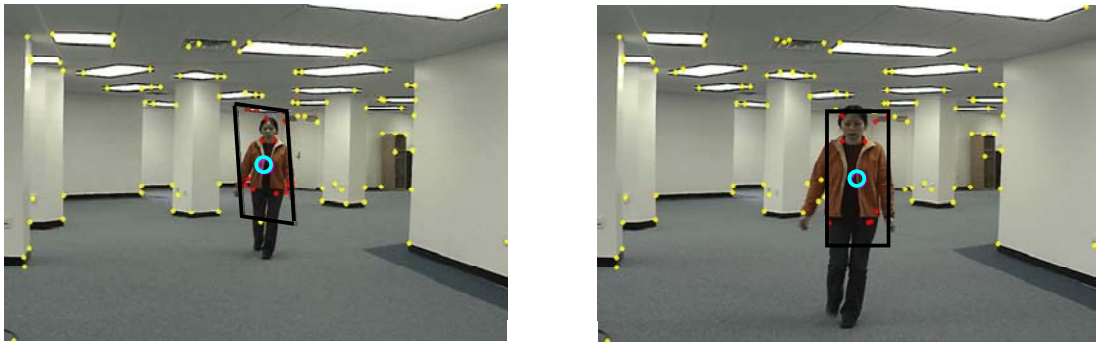
When the target is at a reasonable distance, the algorithms based on both affine and perspective projection models can produce accurate estimation. As the target approaches the camera, the affine projection model is unable to capture the characteristics of the imaging process. The advantage of using the perspective projection model emerges. It can produce accurate estimation regardless of the target's position. The performance of both affine projection models (weak perspective and paraperspective) begins to degrade when the target is at a distance of about 7m. To quantitatively compare their performance, the root mean squared error (RMSE) for the perspective, paraperspective, and weak perspective projection models are computed from the 380^{th} frame when the target is at a distance of approximately 7m and are listed as follows: 0.19, 0.36, and 0.54. As expected, the perspective projection model yields the best accuracy, followed by the paraperspective and weak perspective projection models. However, the performance of the algorithm based on the perspective projection model deteriorates and fails to preserve the necessary robustness when more realistic sequences (deformations, disturbances from background, camera motion) are used.

Since the target's image is close to the image center in this pedestrian sequence, the performances are similar for both affine projection model based algorithms. As the target approaches the camera resulting in increased influence from the center offset, an improved accuracy, quantified by a decreased RMSE from 0.54 to 0.36 and a decreased relative error from 28.6% to 20.0%, is observed from the weak perspective to the paraperspective projection model.

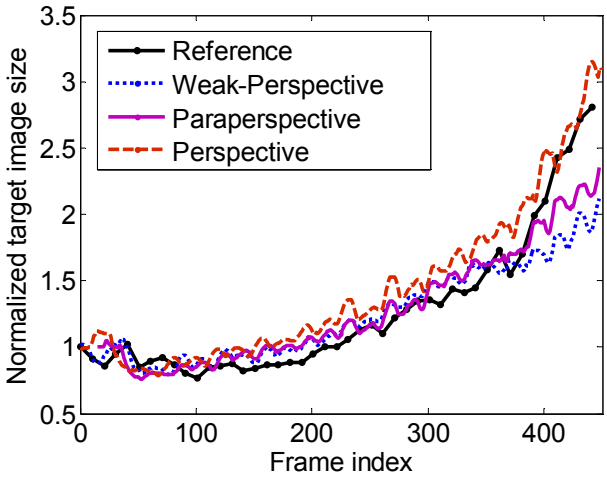
The advantage of using the paraperspective projection model becomes evident when there exists a decent amount of center offset, such as the case in real-time tracking. To manifest the advantage of using the paraperspective projection model, a sequence with center offset is deliberately collected and is shown in Figure 4.3. From Figure 4.3(e), an obvious improvement in estimation accuracy is achieved, indicated by a decrease in the RMSE from 0.31 to 0.12. Since the estimation error is cumulative, more enhanced accuracy is observed from the last frame, where the relative errors for the weak perspective and paraperspective projection models are 17.4% and 3.3%, respectively.



(a) (b)

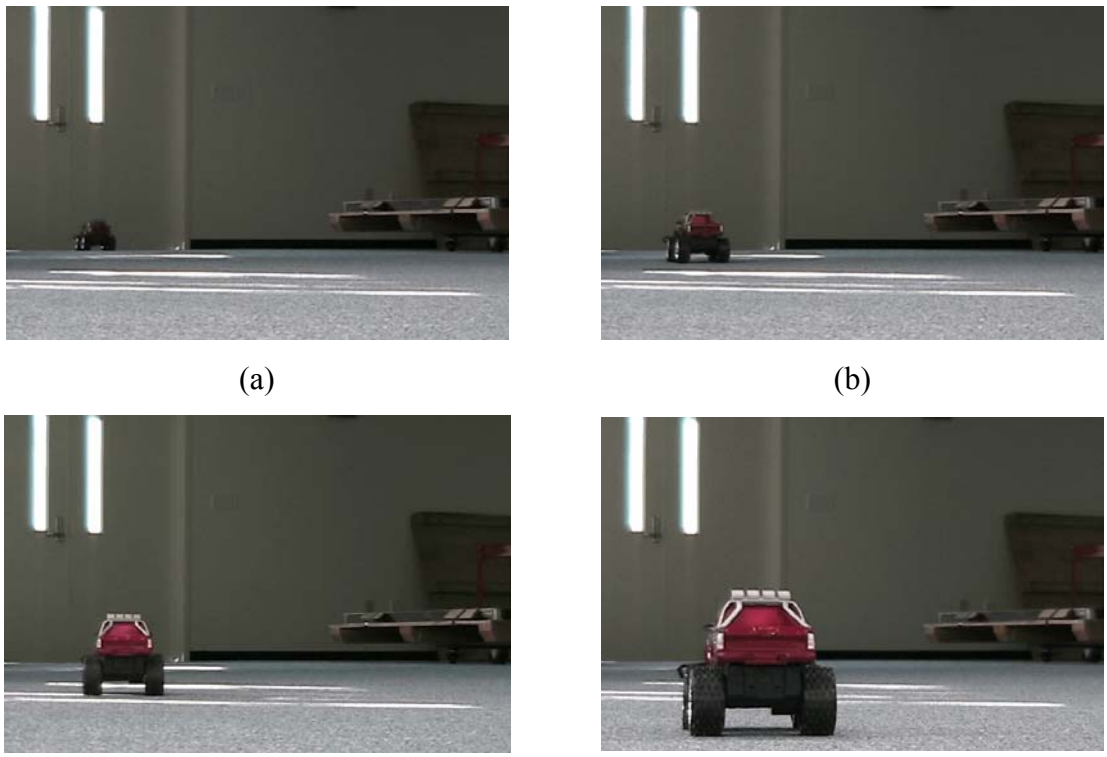


(c) (d)



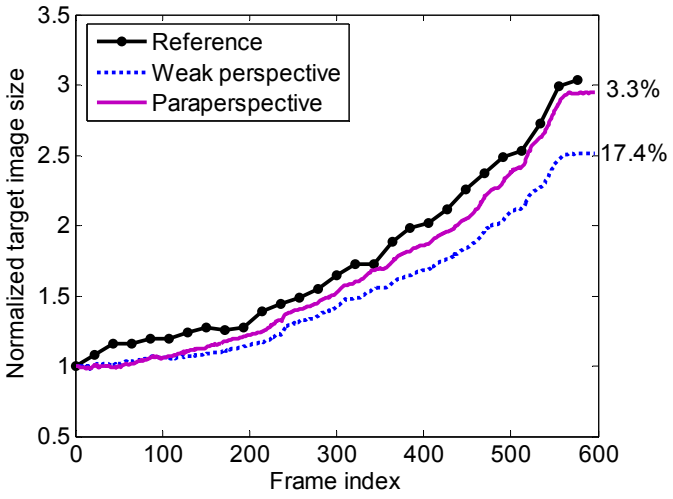
(e)

Figure 4.2. (a)-(d) Sample frames from the pedestrian sequence. Yellow and red dots present the detected corners in the background and foreground, respectively. The affine shape separating the foreground and background corners is depicted by a black quadrilateral. A light blue circle shows the gaze point, to which the camera is directed. (e) Comparison of measured and estimated target scale (normalized to the target's image size in the first frame).



(a) (b)

(c) (d)



(e)

Figure 4.3. (a)-(d) Sample frames from the toy car sequence with center offset. (e) Comparison of measured and estimated target scale (normalized to the target's image size in the first frame).

4.3 Foreground / background segmentation

4.3.1 Theoretical derivation

Given N_a points on the boundaries of the target's image $\mathbf{v}_{i',j} = \begin{bmatrix} v_{x,j}^{i'} & v_{y,j}^{i'} \end{bmatrix}^T$ with $i' = i-1, i-2, \cdots, i-I+1$ and the recovered motion $\begin{bmatrix} M_{i-I+1} \\ \vdots \\ M_i \end{bmatrix}$, we have:

$$\mathbf{v}_{i,j} = M_i \begin{bmatrix} M_{i-I+1} \\ \vdots \\ M_{i-1} \end{bmatrix}^+ \begin{bmatrix} \mathbf{v}_{i-I+1,j} - \mathbf{m}_{i-I+1} \\ \vdots \\ \mathbf{v}_{i-1,j} - \mathbf{m}_{i-1} \end{bmatrix} + \mathbf{m}_i , \tag{4.10}$$

where $(\circ)^+$ denotes the matrix pseudo inverse.

The above affine shape is not stable due to image noise. Without additional constraints, the vertices on the affine shape in the recovered affine space $V_{i,j} = \begin{bmatrix} V_{X,j}^i & V_{Y,j}^i & V_{Z,j}^i \end{bmatrix}^T$ may assume different $V_{Z,j}^i$ and the differences may be exaggerated by the numerical errors in SFM. To avoid unnecessary distortions caused by $V_{Z,j}^i$, it is necessary to impose additional constraints on $V_{Z,j}^i$. Keeping $V_{Z,j}^i$ constant with an unknown value is one possible solution. This is done by constructing:

$$\widetilde{V}_i = \begin{bmatrix} V_{X,1}^i & V_{Y,1}^i & V_{X,2}^i & V_{Y,2}^i & \cdots & \cdots & V_{X,N_a}^i & V_{Y,N_a}^i & V_Z^i \end{bmatrix}^T , \tag{4.11}$$

$$\widetilde{\mathbf{v}}_{i'} = \begin{bmatrix} v_{x,1}^{i'} - m_x^{i'} & v_{y,1}^{i'} - m_y^{i'} & v_{x,2}^{i'} - m_x^{i'} & v_{y,2}^{i'} - m_y^{i'} & \cdots & \cdots & v_{x,N_a}^{i'} - m_x^{i'} & v_{y,N_a}^{i'} - m_y^{i'} \end{bmatrix}^T , \tag{4.12}$$

and

$$\widetilde{M}_{i'} = \begin{bmatrix} m_{11}^{i'} & m_{12}^{i'} & \cdots & 0 & 0 & m_{13}^{i'} \\ m_{21}^{i'} & m_{22}^{i'} & \cdots & 0 & 0 & m_{23}^{i'} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & m_{11}^{i'} & m_{12}^{i'} & m_{13}^{i'} \\ 0 & 0 & \cdots & m_{21}^{i'} & m_{22}^{i'} & m_{23}^{i'} \end{bmatrix} . \tag{4.13}$$

Finally we have:

$$\widetilde{V}_i = \begin{bmatrix} \widetilde{M}_{i-I+1} \\ \vdots \\ \widetilde{M}_{i-1} \end{bmatrix}^{+} \begin{bmatrix} \widetilde{\mathbf{v}}_{i-I+1} \\ \vdots \\ \widetilde{\mathbf{v}}_{i-1} \end{bmatrix}. \tag{4.14}$$

The estimated vertices are $\mathbf{v}_{i,j} = M_i V_{i,j} + \mathbf{m}_i$ with $V_{i,j} = \begin{bmatrix} V_{X,j}^i & V_{Y,j}^i & V_Z^i \end{bmatrix}^T$.

Compared with $M_{i'}$, $\widetilde{M}_{i'}$ has a higher dimension. To save on computations, we further assign $V_{Z,j}^{i'}$ to a fixed and known value, for instance the Z coordinate of the gaze point, $G_Z^{i'}$. Let $M_{i'} = \begin{bmatrix} M_{i'}^{'} \mid \mathbf{m}_{i'}^{'} \end{bmatrix}$, where $M_{i'}^{'}$ includes the first two columns of $M_{i'}$ and $\mathbf{m}_{i'}^{'}$ the third column, respectively. The vertices of the affine shape can be updated by:

$$\mathbf{v}_{i,j} = M_i^{'} \begin{bmatrix} M_{i-I+1}^{'} \\ \vdots \\ M_{i-1}^{'} \end{bmatrix}^{+} \begin{bmatrix} \mathbf{v}_{i-I+1,j} - \mathbf{m}_{i-I+1} - G_Z^{i-I+1}\mathbf{m}'_{i-I+1} \\ \vdots \\ \mathbf{v}_{i-1,j} - \mathbf{m}_{i-1} - G_Z^{i-1}\mathbf{m}'_{i-1} \end{bmatrix} + G_Z^i \mathbf{m}'_i + \mathbf{m}_i . \tag{4.15}$$

Since the origin of the reconstructed affine basis corresponds to the center of mass of all tracked corners, we can also set $V_{Z,j}^{i'} = 0$ and arrive at:

$$\mathbf{v}_{i,j} = M_i^{'} \begin{bmatrix} M_{i-I+1}^{'} \\ \vdots \\ M_{i-1}^{'} \end{bmatrix}^{+} \begin{bmatrix} \mathbf{v}_{i-I+1,j} - \mathbf{m}_{i-I+1} \\ \vdots \\ \mathbf{v}_{i-1,j} - \mathbf{m}_{i-1} \end{bmatrix} + \mathbf{m}_i . \tag{4.16}$$

In so doing, no prior solution of $G_Z^{i'}$ is necessary.

The Z axis of the recovered affine space corresponds to the direction with the third largest singular value of the measurement matrix W_i in (4.3), equivalently the smallest dimension in the target's reconstructed geometry or the smallest variations in the target's relative motion. By choosing a fixed $V_{Z,j}^i$, we only consider the variations in the first two principal axes. By forcing $V_{Z,j}^{i'} = G_Z^{i'}$ or $V_{Z,j}^{i'} = 0$, the affine shape is centered at the gaze point or the target's center of mass. This arrangement is representative of the target's 3D geometry and motion.

The aforementioned algorithm is efficient in handling general motions including off-plane rotation. However, as the target rotates, the affine shape keeps tracking and rotating with the originally visible sides but excludes the newly detected target corners in the previously hidden sides. When the rotation angle is large, the hidden sides of the object become dominant while the originally visible sides diminish. With fewer and fewer matched features, the system performance deteriorates.

To be able to include the newly detected corners in the previously hidden sides and accommodate a large off-plane rotation angle, the variations along the Z axis must be taken into consideration as well. Therefore, two affine shapes are used (Figure 4.4), each passing the extreme points of the Z axis and parallel to the plane determined by the other

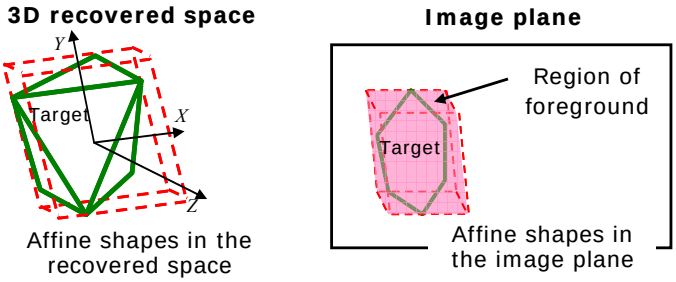


Figure 4.4. Illustration of the use of two affine shapes.

two affine basis, forming a 3D affine shape. In so doing, the relative position between these two affine shapes embodies the changes in the Z axis.

In summary, the proposed foreground/background segmentation algorithm proceeds as follows.

1. Initialize the affine shape.
   1.1 Obtain the reconstructed structure $P_{0,j}$ from the SVD factorization of the measurement matrix W_0.
   1.2 Find the maximum $Z^0_{\max} = \max_j \{Z_{0,j}\}$ and minimum coordinates $Z^0_{\min} = \min_j \{Z_{0,j}\}$ along the third affine axis.
   1.3 Construct two affine shapes V_0^- and V_0^+ based on these extreme points:

$$V_0^- = \begin{bmatrix} V_{0,1}^- & \cdots & V_{0,N_a}^- \end{bmatrix} = \begin{bmatrix} V_{X,1}^0 & \cdots & V_{X,N_a}^0 \\ V_{Y,1}^0 & \cdots & V_{Y,N_a}^0 \\ Z_{\min}^0 & \cdots & Z_{\min}^0 \end{bmatrix},$$

   and

$$V_0^+ = \begin{bmatrix} V_{0,1}^+ & \cdots & V_{0,N_a}^+ \end{bmatrix} = \begin{bmatrix} V_{X,1}^0 & \cdots & V_{X,N_a}^0 \\ V_{Y,1}^0 & \cdots & V_{Y,N_a}^0 \\ Z_{\max}^0 & \cdots & Z_{\max}^0 \end{bmatrix}.$$

   1.4 Project the affine shapes onto the 2D image plane: $\mathbf{v}_{0,j}^{\pm} = M_0 V_{0,j}^{\pm} + \mathbf{m}_0$.
2. Update the affine shapes by:

$$\mathbf{v}_{i,j}^- = M_i' \begin{bmatrix} M_{i-I+1}' \\ \vdots \\ M_{i-1}' \end{bmatrix}^+ \begin{bmatrix} \mathbf{v}_{i-I+1,j}^- - \mathbf{m}_{i-I+1} - Z_{\min}^{i-I+1} \mathbf{m}'_{i-I+1} \\ \vdots \\ \mathbf{v}_{i-1,j}^- - \mathbf{m}_{i-1} - Z_{\min}^{i-1} \mathbf{m}'_{i-1} \end{bmatrix} + Z_{\min}^i \mathbf{m}'_i + \mathbf{m}_i,$$

   and

$$\mathbf{v}_{i,j}^+ = M_i' \begin{bmatrix} M_{i-I+1}' \\ \vdots \\ M_{i-1}' \end{bmatrix}^+ \begin{bmatrix} \mathbf{v}_{i-I+1,j}^+ - \mathbf{m}_{i-I+1} - Z_{\max}^{i-I+1} \mathbf{m}'_{i-I+1} \\ \vdots \\ \mathbf{v}_{i-1,j}^+ - \mathbf{m}_{i-1} - Z_{\max}^{i-1} \mathbf{m}'_{i-1} \end{bmatrix} + Z_{\max}^i \mathbf{m}'_i + \mathbf{m}_i.$$

3. Construct the region of foreground *ROF* with $\mathbf{v}_{i,j}^-$ and $\mathbf{v}_{i,j}^+$ as the vertices in the

$(i+1)^{th}$ frame.

4. Separate the foreground corners in the $(i+1)^{th}$ frame from background if $\mathbf{p}_{i+1,j} \in ROF$.

The use of two affine shapes or equivalently one 3D affine shape helps to capture the target's motion more precisely. The changes caused by the target's off-plane rotation can be incorporated as well. A quadrangle affine shape with $N_a = 4$ is used in our experiments. More points can be added producing general shapes such as polygons and ellipses. An extreme case is to use the points on the target's detected contour, where the affine shape and an active contour algorithm can be applied to predict and refine the vertices, respectively. The choice of the shapes is application dependent.

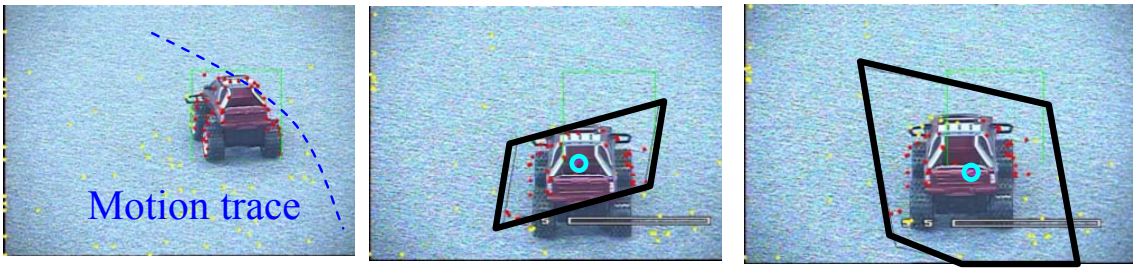
The prominent advantage of the proposed segmentation method is its fast implementation and low computational cost. Since the affine structure is readily computed for the purpose of scale estimation, the additional computation is only an update for N_a points. Once the affine shape is determined, the separation reduces to identifying the corners inside the *ROF*. To improve the segmentation accuracy, multiple cues can be explored to reject outliers. Color information is a popular choice. After the first round selection by the affine shape, the number of candidate points is much smaller. Thus, the successive process of rejecting outliers can be carried out without bringing in noticeably increased complexity.

The ability to handle off-plane rotation is also nontrivial. The target presents a different view to the camera during rotation, which impedes the use of appearance based methods [Collins03, Kim03] unless a timely update is conducted. Although RANSAC based algorithms can be used at the cost of considerably increased computations [Tordoff04], it is equally difficult to separate foreground features while the target is rotating. The ability of rotation handling, provided by the incorporation of the target's 3D geometry and motion, facilitates view independent target pursuing and provides a promising segmentation method when off-plane rotation is involved.

The effectiveness of our foreground and background segmentation depends on the accuracy of the reconstructed structure, which in turn relies on the efficiency of SFM. The accuracy of the segmentation is dominated by the percentage, relative position, and relative motion of the erroneously classified foreground corners with respect to the correctly classified corners. Our algorithm is able to produce accurate and robust segmentation if the majority of the matched corners are correctly recognized. From our experiments, our algorithm has a tolerance of 20% for erroneously classified corners. The amount of the tolerance is obtained empirically by observing the accuracy of object tracking and scale estimation after purposefully including points close to the target but lying outside the estimated affine shape. This tolerance is sufficient for most practical surveillance systems.

4.3.2 Experimental results

The affine shape with unconstrained $V_{Z,j}^{i}$ is not stable and may undergo sudden distortions even under controlled scenarios such as the toy car sequence (Figure 4.5). In



(a) (b) (c)

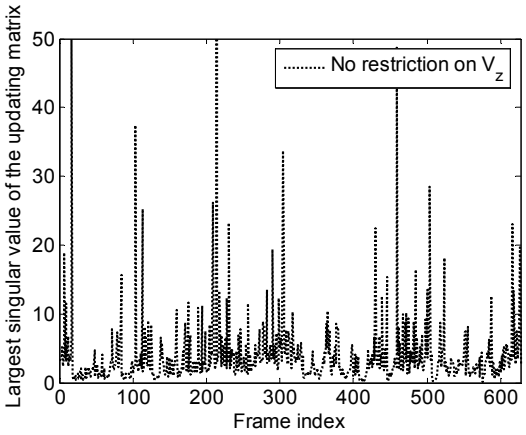
Figure 4.5. Sample frames from the toy car sequence with unconstrained $V_{Z,j}^{i}$. (a) Reference frame. (b) and (c) Two consecutive frames with a sudden change in the estimated affine shape.

this sequence, the target moves toward the camera at a constant speed from a distance of 10m to a distance of 5m and rotates approximately 20°. Two consecutive frames are shown in Figures 4.5(b) and (c), where a sudden distortion in the estimated affine shape occurs.

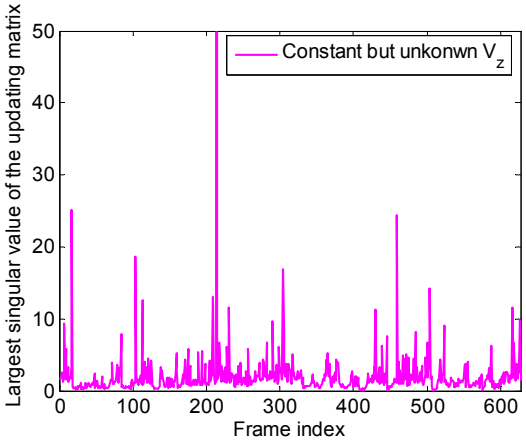
To fully understand the cause of the observed sudden changes and illustrate the effect of restricting $V_{Z,j}^{i}$, we look into the behavior of the updating matrix $M_i^+ = \begin{bmatrix} M_{i-I+1} \\ \vdots \\ M_{i-1} \end{bmatrix}^+$ or $\tilde{M}_i^+ = \begin{bmatrix} \tilde{M}_{i-I+1} \\ \vdots \\ \tilde{M}_{i-1} \end{bmatrix}^+$ for cases with constant $V_{Z,j}^{i}$. The largest singular value of M_i^+, $S_{M_i^+}$, is studied since it best describes the characteristics of M_i^+. From Figure 4.6(a), we observe frequent spikes in the plot of S_{M^+}. These spikes are responsible for the sudden distortions in the affine shape.

From unconstrained $V_{Z,j}^{i}$ to constant but unknown $V_{Z,j}^{i}$ as shown in Figure 4.6(b), although the spikes occur occasionally, the variations in S_{M^+} are reduced, with the standard deviation decreasing from 11.74 to 5.87. In Figure 4.6(c), $V_{Z,j}^{i}$ is restricted to a fixed and known value. The observed variations further decrease with the standard deviation dropping from 5.87 to 1.35. More importantly, there are no visible spikes in S_{M^+}, eliminating undesired distortions in the affine shape.

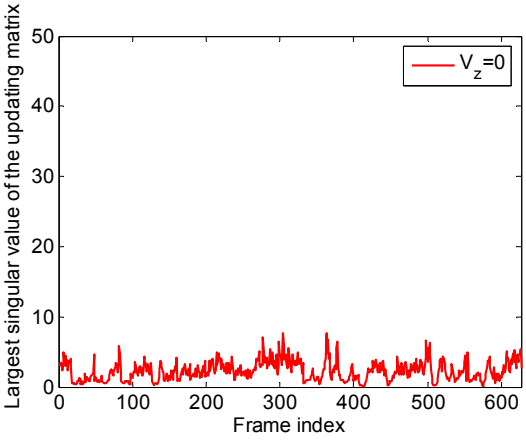
Under the similar experimental condition, the affine shape with $V_{Z,j}^{i} = 0$ presents considerably improved stability, as shown in Figure 4.7. The resulting affine shape is able to trace the moving target closely and maintain its shape consistently throughout the sequence. Considering stability and computational complexity, the affine shape with a



(a)

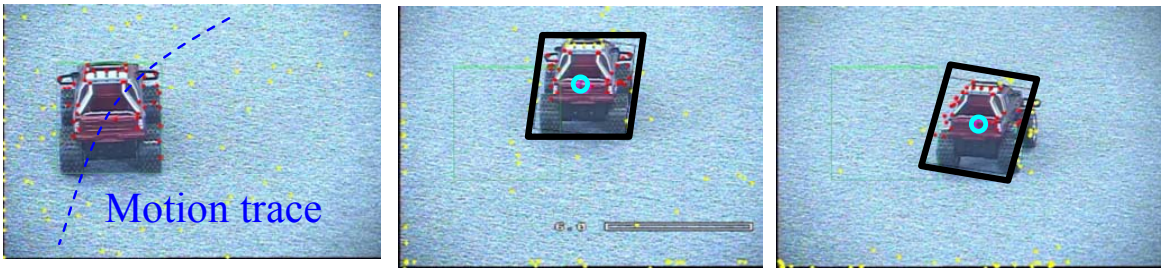


(b)



(c)

Figure 4.6. Comparison of $S_{M_i^+}$ based on the toy car sequence with various $V_{Z,j}^i$: (a) unconstrained $V_{Z,j}^i$, (b) constant but unknown $V_{Z,j}^i$, and (c) $V_{Z,j}^i = 0$. The values of $S_{M_i^+}$ become increasingly stable from unconstrained $V_{Z,j}^i$ to $V_{Z,j}^i = 0$, which eliminates distortions in the affine shapes.



(a) (b) (c)

Figure 4.7. Sample frames from the toy car sequence with $V_{Z,j}^{i} = 0$. (a) Reference frame. (b) and (c) Frames before and after rotation. The affine shape follows the target closely with no distortions as the target rotates.

constant and known $V_{Z,j}^{i}$ produces the best performance. From our experiments the actual value of $V_{Z,j}^{i}$, either $V_{Z,j}^{i} = G_{Z}^{i}$ or $V_{Z,j}^{i} = 0$, does not affect the algorithm's stability.

The following experiment examines the performance of the proposed 3D affine shape in handling a large amount of rotation (>45°). The subject remains at the same position and turns his head from 0° to 90° and then backwards. Figure 4.8 demonstrates sample frames with a single affine shape implemented. The resulting affine shape is able to locate and trace the originally visible side of the target (the frontal view of the target's face) precisely. However, apparently, with a single affine shape, the variations along the *Z* axis are absent and the resulting affine shape closely resembles the target's frontal view. As a result, the newly emerged features from the target's side view, to be more specific the features near the target's ear, are excluded.

Figure 4.9 shows sample frames with two affine shapes or equivalently one 3D affine shape implemented. It is obvious that our algorithm is now capable of handling large degrees of deformation and accommodating rotation. Compared with Figures 4.8(b) and (c), features from the target's side view are included automatically.

Figure 4.10 illustrates sample frames when the toy car moves toward the camera from a distance of 10m to a distance of 3m. Figure 4.10(e) shows the estimated target scale with automatic zoom control and compares it with the scale change if the camera's zoom is kept constant. With automatic zoom control where the camera's zoom is varied from 9× to 3× approximately, the target's image size is maintained with a variation of 10% of the original scale. Figures 4.11 and 4.12 demonstrate sample frames and estimated target scale from real-time pedestrian sequences. In Figure 4.11, the camera's zoom is changed from 8× to 2× so that the target's image size is maintained with only slight variations. In parallel, in Figure 4.12, the camera's zoom is changed from 4× to 1× automatically resulting in a relative variation less than 6% of the total variations if no size preserving zoom control is applied.

(a) (b)

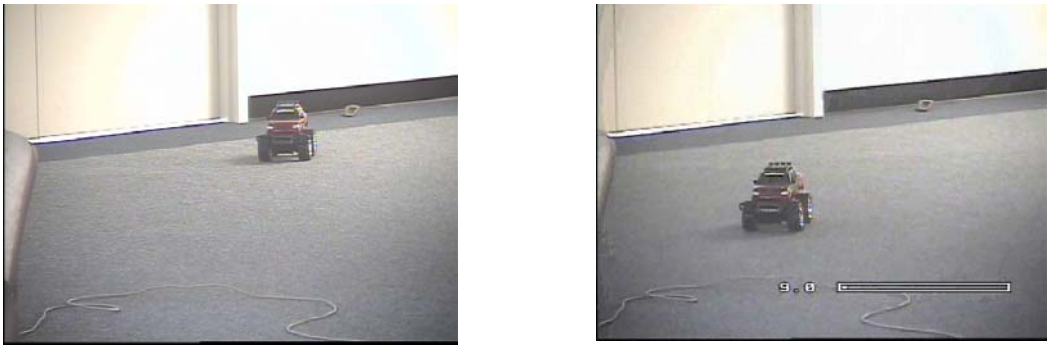
(c) (d)

Figure 4.8. Sample frames from the men's face sequence with a single affine shape ($V^i_{Z,j} = 0$). (a) 0°. (b) 90°. (c) 45°. (d) 0°. The estimated affine shape follows the frontal view of the target, which is initially visible. The newly detected points on the initially invisible views (side view) are not considered as the foreground.

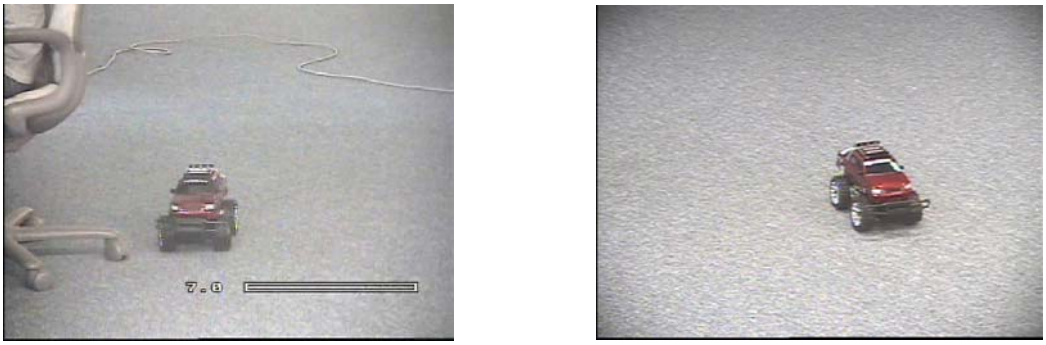
(a) (b)

(c) (d)

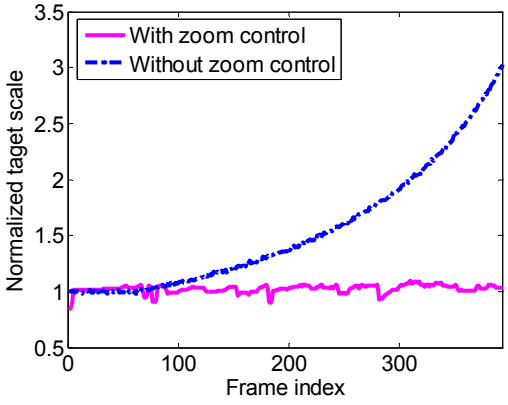
Figure 4.9. Sample frames from the men's face sequence with two affine shapes depicted in blue and green. (a) 0°. (b) 90°. (c) 45°. (d) 0°. The use of two affine shapes ensures that the newly detected points on the initially invisible views (side view) are automatically considered as the foreground.



(a) (b)



(c) (d)



(e)

Figure 4.10. (a)-(d) Sample frames from a real-time toy car sequence. (e) Estimated target scale (normalized to the target's image size in the first frame).

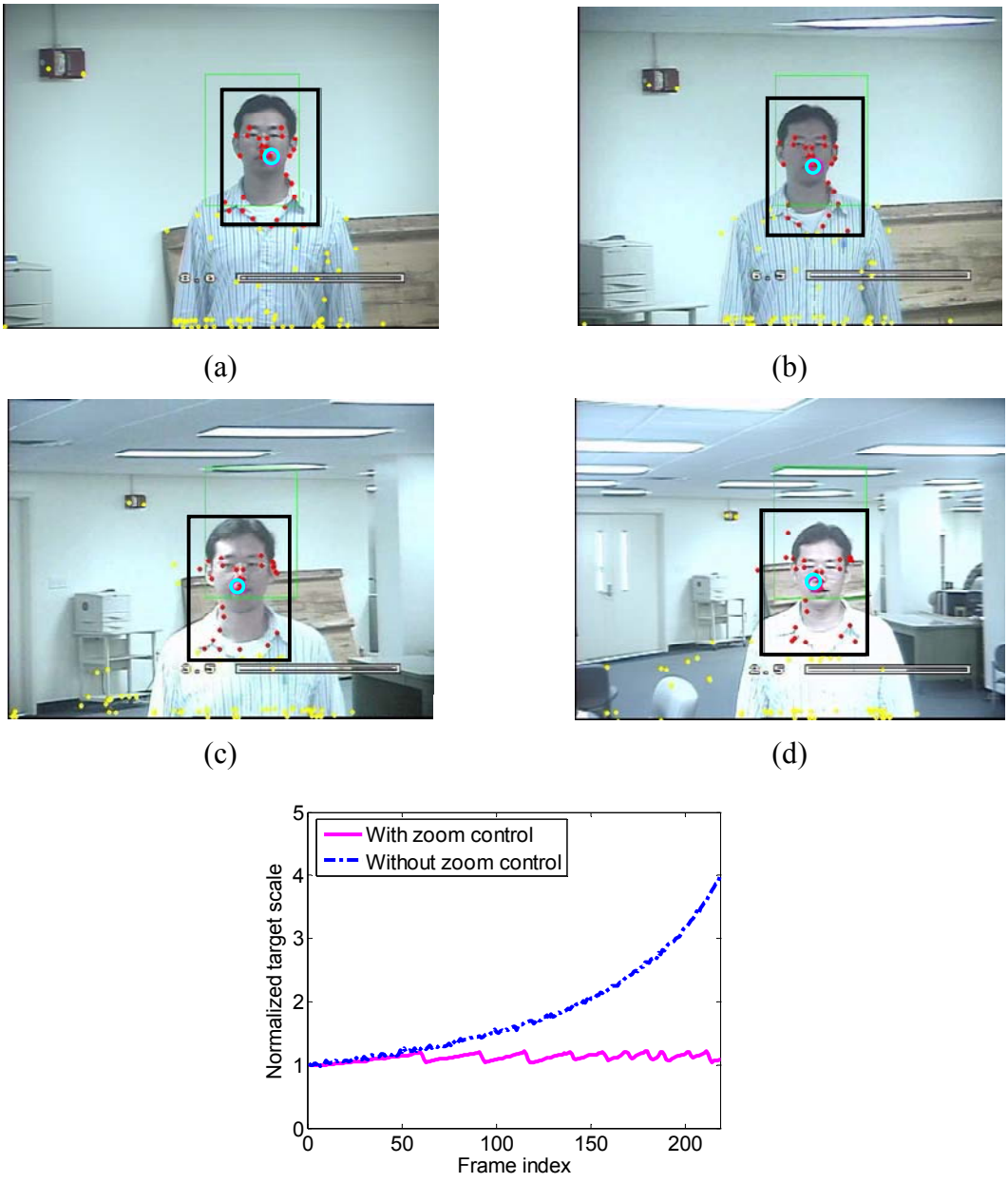
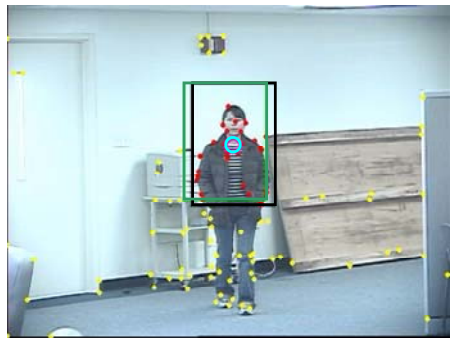
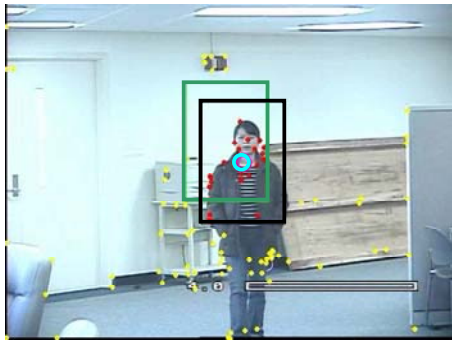


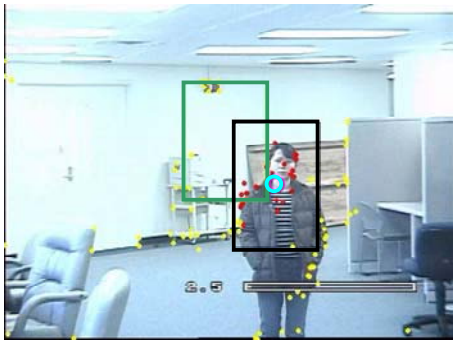
Figure 4.11. (a)-(d) Sample frames from a real-time pedestrian sequence including busy background, illumination change, and pose variation. Green bounding box illustrates the target's initial image size, which is to be preserved throughout the sequence. (e) Estimated target scale (normalized to the target's image size in the first frame). The face resolution is maintained throughout the sequence.



(a)

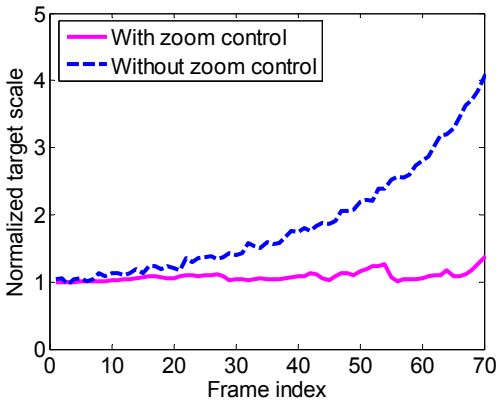


(b)



(c)

(d)



(e)

Figure 4.12. (a)-(d) Sample frames of a real-time pedestrian sequence including busy background and illumination change. (e) Estimated target scale (normalized to the target's image size in the first frame). The target scale is maintained throughout the sequence. Note that due to system latency, there exists a moderate amount of center offset in this sequence.

5 Camera handoff

Camera handoff automatically governs the collaboration among adjacent cameras in real time to ensure persistent surveillance. As the first step of camera handoff, the observation measures of the tracked targets are computed and used to determine when to trigger a handoff request. Once a handoff request is issued, the adjacent cameras' computational load is examined to select available candidate cameras. These candidate cameras then exchange information and perform consistent labeling. A hybrid consistent labeling algorithm is employed in our system, where both geometry relation and color information are used for data association among cameras. Finally, the tracked target is transferred to the optimal candidate camera considering resolution, occlusion, camera load, frontal view, *etc*.

The definition of the observation measure used in camera handoff is given in section 5.1. Section 5.2 presents our camera handoff algorithm. Experimental results are demonstrated in section 5.3.

5.1 Observation measure

The observation measure discussed in section 3.1 is designed for sensor planning, where a full knowledge of the camera's intrinsic and extrinsic parameters is assumed. In practical surveillance systems, this assumption needs to be relaxed. In this section we design the observation measure for camera handoff purely based on the 2D input sequences. The definitions given herewith approximate the counterparts in section 3.1. The requirements on designing the observation measure for camera handoff are two folded. (1) The observation measure for camera handoff should represent the corresponding quantities defined for sensor planning, which establishes the connection between camera handoff and sensor planning. This is important since it ensures that the optimal handoff success rate as predicted by sensor planning can be achieved in practical surveillance. (2) The computations of the observation measure for camera handoff only depend on quantities derivable from 2D images with no prior knowledge of the camera's intrinsic and extrinsic parameters. Linear and direct computations are preferred for real-time applications. In parallel to the layout of section 3.1, the observation measure for camera handoff is defined for static perspective, PTZ, and omnidirectional cameras.

5.1.1 Static perspective cameras

From the size preserving tracking algorithm discussed in chapter 4, the target scale is readily estimated. Therefore, the resolution component M_R is:

$$M_R = \alpha_R \, f / Z_r = \alpha_R \rho \,. \tag{5.1}$$

The gaze point $\mathbf{g} = \begin{bmatrix} g_x & g_y \end{bmatrix}^T$ is used to evaluate the distance to the edges of the camera's FOV:

$$M_D = \left\{ \alpha_D \left[\left(1 - \frac{|\, g_x \,|}{N_{row}/2} \right)^2 + \left(1 - \frac{|\, g_y \,|}{N_{col}/2} \right)^2 \right] \right\}^{\beta_1 \rho + \beta_0} . \tag{5.2}$$

From the estimated gaze point, we can also compute its velocity $\mathbf{u}_g$. It is sufficient to use the projected velocity along the camera's optical axis $u_{g,x}$ to describe the frontal view component:

$$M_{FV} = \alpha_{FV} \frac{-u_{g,x}}{\left\| \mathbf{u}_g \right\|} . \tag{5.3}$$

Since 3D reconstruction is also obtained based on the affine projection model in our size preserving tracking, the target's 3D trace can be estimated. We can also follow the definition in (3.10) for a more accurate estimation of M_{FV}. However, from our experiments, the above definition produces an acceptable accuracy with considerably decreased computational complexity. The observation measure for camera handoff is a weighted sum of these three components:

$$Q = \begin{cases} w_R M_R + w_D M_D + w_{FV} M_{FV} & [x \;\; y]^T \in \Pi \\ -\infty & otherwise \end{cases} . \tag{5.4}$$

5.1.2 PTZ cameras

Under the assumption that the object of interest is maintained within the proximity of the image center due to proper pan and tilt controls, the M_D component for PTZ cameras remains approximately constant independently from the target's relative position and, therefore, can be omitted. The resolution and frontal view components follow the definitions for static perspective cameras:

$$M_R = \alpha_R \, f / Z_r = \alpha_R \rho \,, \tag{5.5}$$

and

$$M_{FV} = \alpha_{FV} \frac{-u_{g,x}}{\|\mathbf{u}_g\|} . \tag{5.6}$$

Note that different from static perspective cameras the focal length in (5.5) is indeed time varying for PTZ cameras. Our size preserving tracking algorithm discussed in chapter 4 considers the changes in target scale caused by a time varying focal length as well and outputs an estimate of f/Z_r including the combined effect of camera zooming and target motion.

5.1.3 Omnidirectional cameras

From section 3.1, the resolution component for an omnidirectional camera is defined as:

$$M_R = \frac{\alpha_R f Z}{Z^2 + R^2} \sum_{k=1,odd} \lambda_{\theta,k} k \theta^{k-1} . \tag{5.7}$$

From the centroid of the detected motion blob, $\mathbf{g} = \begin{bmatrix} g_x & g_y \end{bmatrix}^T$, the angle between the incoming ray and the camera's optical axis θ can be solved by computing the roots of $r = f \sum_{k=1,odd} \lambda_{\theta,k} \theta^k$, with $\theta \in [0, \frac{\pi}{2})$ and $r = \sqrt{g_x^2 + g_y^2}$. Given the estimated $\hat{\theta}$, the M_R component is expressed as:

$$M_R = \frac{\alpha_R f \cos^2 \hat{\theta}}{Z} \sum_{k=1,odd} \lambda_{\theta,k} k \hat{\theta}^{k-1} . \tag{5.8}$$

The maximum of M_R is achieved at $\hat{\theta} = 0$, which yields $M_{R,\max} = \frac{\alpha_R f \lambda_{\theta,1}}{Z}$. To normalize the M_R component between zero and one, we have:

$$M_R = \frac{\cos^2 \hat{\theta}}{\lambda_{\theta,1}} \sum_{k=1,odd} \lambda_{\theta,k} k \hat{\theta}^{k-1} . \tag{5.9}$$

The M_D component is given by:

$$M_D = \alpha_D \left(1 - r / r_o\right)^2 , \tag{5.10}$$

and the definition of the frontal view component remains the same:

$$M_{FV} = \alpha_{FV} \frac{-u_{g,x}}{\|\mathbf{u}_g\|}. \tag{5.11}$$

5.2 Algorithm description

Figure 5.1 illustrates the flow chart of our camera handoff algorithm, where operations are carried at the handoff request and handoff response ends. Let the j^{th} camera be the handoff request end and the i^{th} target be the one that needs a transfer. A handoff request is triggered and broadcasted if $Q_{ij} < Q_T$ and Q_{ij} is decreasing. Afterwards, the j^{th} camera keeps tracking the i^{th} target and waits for confirmation responses from adjacent cameras while the target is still visible.

At the handoff response end, the $(j')^{th}$ camera examines its current computational load and rejects the handoff request if the maximum number of objects that can be tracked simultaneously $N_{obj,j'}$ has been achieved. Otherwise, the response camera checks the probability of the i^{th} target being occluded by other tracked targets in its FOV. A positive response is granted if the probability of dynamic occlusion is low. If the i^{th} target falls in the dynamic occlusion of the $(i')^{th}$ target, their observation measures are compared. The handoff request is rejected if $Q_{ij'} \le Q_{i'j'}$. Otherwise, a handoff request for the $(i')^{th}$ target is triggered by the $(j')^{th}$ camera. Meanwhile, a positive handoff response for the i^{th} target is issued.

Back to the handoff request end, if no confirmation response is received before the j^{th} camera loses track of the i^{th} target, a handoff failure is issued. Otherwise, among all available candidate cameras, the one with the highest observation measure is chosen $j^* = \arg\max\{Q_{ij'}\}$ and the i^{th} target is transferred from the j^{th} camera to the $(j^*)^{th}$ camera if $Q_{ij} < Q_{ij^*}$ and Q_{ij^*} is increasing.

In practical surveillance, we need to address the influence of noise. Because of non-ideal tracking and consistent labeling, the resulting observation measures are noisy. The rule of selecting the camera with the largest observation measure is not entirely valid. Instead, we want to choose the optimal camera by maximizing the probability of the corresponding observation measure being the maximum among competing cameras.

To begin our discussion, the noise introduced by non-ideal tracking and consistent labeling is assumed to follow a Gaussian distribution. An extended Kalman filter can be constructed based on a state vector $\begin{bmatrix}\rho & \Delta\rho & g_x & g_y & u_{g,x} & u_{g,y}\end{bmatrix}^T$ and a measurement vector $\begin{bmatrix}\rho & g_x & g_y & Q\end{bmatrix}^T$. From the output of the extended Kalman filter, the *a posteriori* probability of the observation measure at the i^{th} grid from the j^{th} camera $p(Q_{ij}, \mu_{Q,ij}, \sigma_{Q,ij})$ follows a Gaussian distribution with mean $\mu_{Q,ij}$ and standard deviation $\sigma_{Q,ij}$ [Grewal01]. We then design the following cost function based on Q_{ij} to govern the transition between cameras:

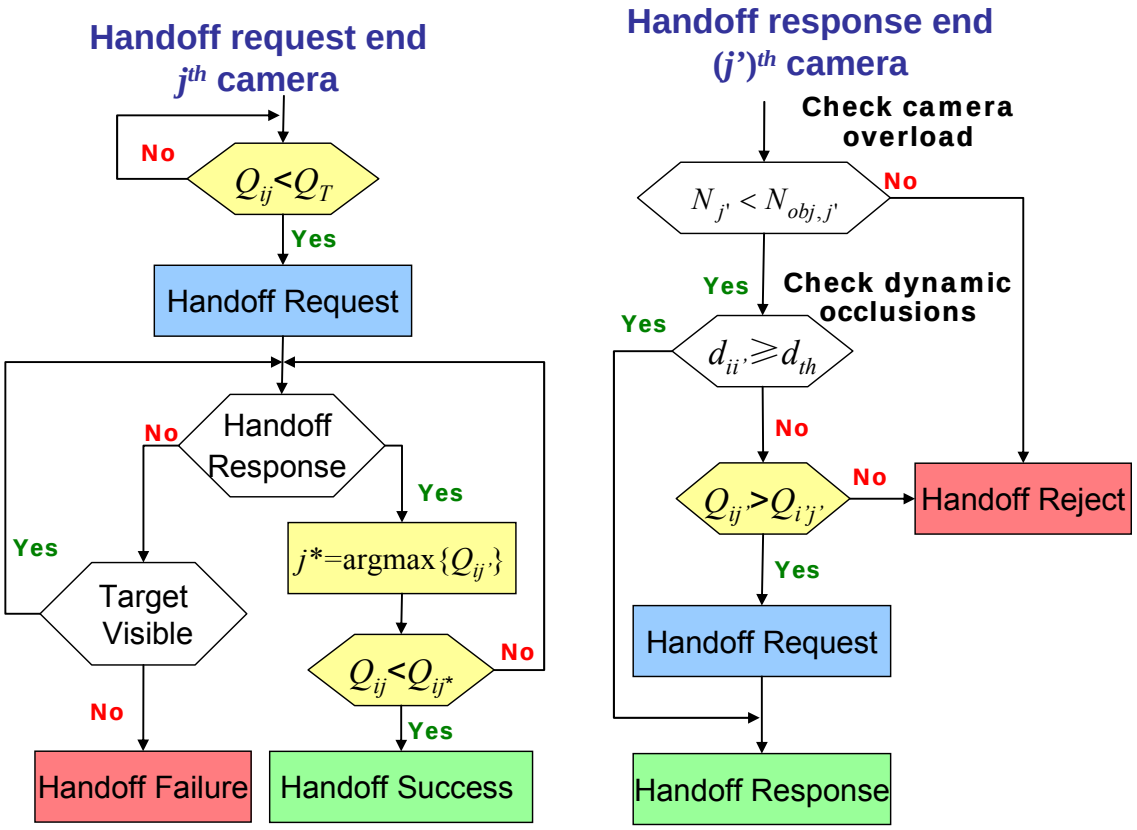


Figure 5.1. Flow chart of the camera handoff algorithm. The handoff response end has the ability to handle dynamic occlusion and camera overload. $N_{j'}$ is the number of tracked objects in the $(j')^{th}$ camera and $N_{obj,j'}$ is the maximum number of objects that can be tracked simultaneously by the $(j')^{th}$ camera. $d_{ii'}$ denotes the distance between the images of the i^{th} and $(i')^{th}$ targets and d_{th} is a predefined threshold for dynamic occlusions.

$$\psi_{ij} = \log\left(p(Q_{ij} > Q_{ij'} for\ j' \neq j)\right) = \sum_{j'=1, j' \neq j}^{N_{cam}} \log\left(\int_{-\infty}^{Q_{ij}} p(Q_{ij'}, \mu_{Q,ij'}, \sigma_{Q,ij'}) dQ_{ij'}\right), \quad (5.12)$$

where N_{cam} denotes the number of candidate cameras. The optimal camera is selected according to $j^* = \arg\max\{\psi_{ij}\}$.

5.3 Experimental results

In this section, we study the effectiveness of the newly defined observation measure in triggering and executing camera handoffs based on real-time pedestrian sequences. Due to the radial symmetric property of omnidirectional cameras, it is sufficient to examine the target's motion across the camera's FOV. As for static perspective cameras, camera handoffs are triggered by the target's motions along and orthogonal to the camera's optical axis. To obtain a statistically valid performance evaluation and comparison between Q_{ij} and ψ_{ij}, simulations based on 300 randomly generated traces are conducted for both static perspective and omnidirectional cameras.

5.3.1 Camera handoff between omnidirectional cameras

In the following experiments, two omnidirectional cameras, IQEye3 and IPIX are used to test our camera handoff algorithm. As shown in Figure 5.2, two omnidirectional cameras are placed at the same height of 3m and at 10m apart. They are calibrated before hand using the algorithm described in [Yao06E]. A polynomial of degree one is selected as the optimal model by the Akaike information criterion for the IQEye3 camera: $r = 1.342\theta$. As for the IPIX camera, polynomials of degree one, $r = 1.4055\theta$, and degree three, $r = 1.4435\theta + 0.0175\theta^3$, present similar performances. For a low computational cost

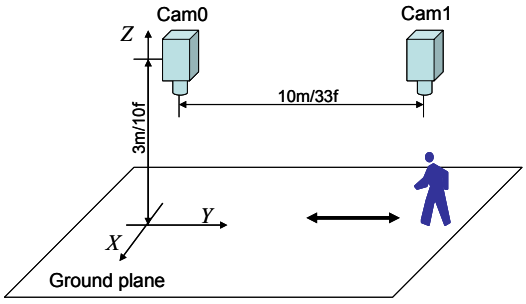


Figure 5.2. Schematic illustration of the system setup for experiments using two omnidirectional cameras.

in camera handoff, the polynomial of degree one is used. From the known relative positions of both cameras, we are able to derive their spatial correspondences under the assumption of planar motion. The experiment in Figure 5.3 illustrates the process of triggering and executing camera handoffs and the experiment in Figure 5.4 investigates the robustness of our handoff algorithm in the presence of dynamic occlusion.

From Figure 5.3, we can see that the target is first detected and tracked by camera 1. As the target moves across the FOV of camera 1, the corresponding observation measure first increases and then decreases. A handoff request is triggered as the target further approaches the edges of the camera's FOV. In the meanwhile, the target also appears in the FOV of camera 0 so that consistent labeling can be established. Afterwards, the target is taken over by camera 0 before it becomes untraceable in camera 1. Similar process repeats as the target moves in the opposite direction and returns to the starting position. The newly defined observation measure accurately describes the quality of object tracking and detects the moment when the tracked target requires a handoff. As a result, camera handoffs are executed successfully and smoothly. The target is tracked continuously and consistently.

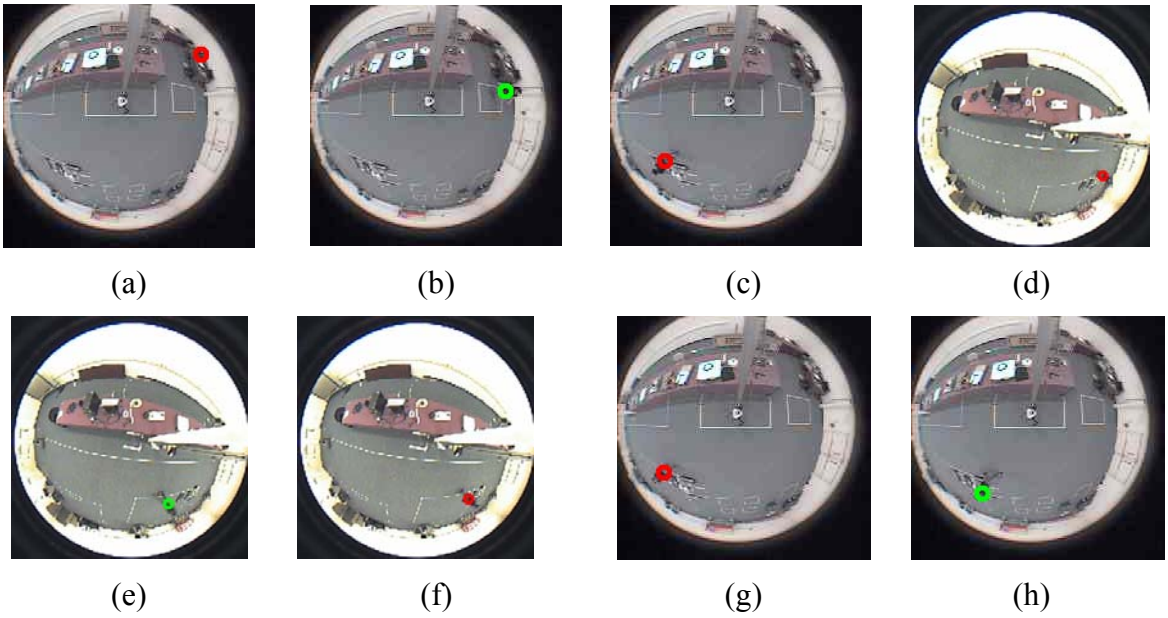
The proposed handoff algorithm is capable of handling partial occlusion, as show in Figure 5.4. Two targets are tracked simultaneously and transferred between two omnidirectional cameras smoothly regardless of partial occlusions.

5.3.2 Camera handoff between static perspective cameras

Two perspective cameras are placed at the same height of 2m. Three scenarios are examined according to the angle between the optical axes of the cameras: 0°, 180°, and 90°. Figure 5.5 schematically illustrates the experimental setups. We refer to these three setups as case A, B, and C.

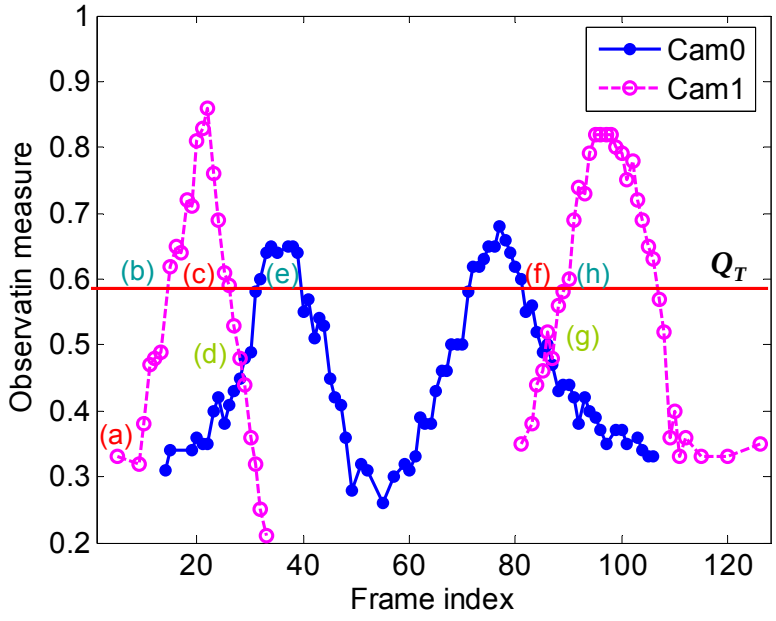
For case A, as the target first walks into the FOV of camera 0, the corresponding observation measure is below the trigger threshold $Q<Q_T$. In Figure 5.6(b) the observation measure achieves $Q \geq Q_T$. As the target moves toward the edges of the camera's FOV, the observation measure falls below the trigger threshold again. Thus, at the position shown in Figure 5.6(c), a handoff request is issued. From Figure 5.6(c) to (d), communications and consistent labeling between camera 0 and camera 1 are established. Figure 5.6(d) is the last frame where the target is tracked by camera 0 and Figure 5.6(e) is the first frame after the target is transferred to camera 1. Afterwards, since there is no additional camera to take it over from camera 1, the target is tracked by camera 1 even after a handoff is triggered in Figure 5.6(f). Camera 1 continues tracking the target until it falls out of the camera's FOV as shown in Figure 5.6(g). Sufficient overlapped FOVs are reserved for camera handoff and the pedestrian is successfully handed over from camera 0 to camera 1.

Figure 5.6(h) depicts the resulting observation measure for both cameras. The distance to the edges of the camera's FOV component dominates the transition between the two cameras since the resolution and frontal view components are kept approximately



(a) (b) (c) (d)

(e) (f) (g) (h)



(i)

Figure 5.3. Camera handoff between omnidirectional cameras. (a) First frame with the detected target in camera 1. (b) Tracked target with $Q \geq Q_T$ in camera 1. (c) Triggered handoff in camera 1. (d) Handoff is executed. The tracked target is transferred from camera 1 to camera 0. (e) Tracked target with $Q \geq Q_T$ in camera 0. (f) Triggered handoff in camera 0. (g) Handoff is executed. The tracked target is transferred from camera 0 to camera 1. (h) Tracked target with $Q \geq Q_T$ in camera 1. (i) Observation measure. The observation measures of frames (a)-(h) are specified.

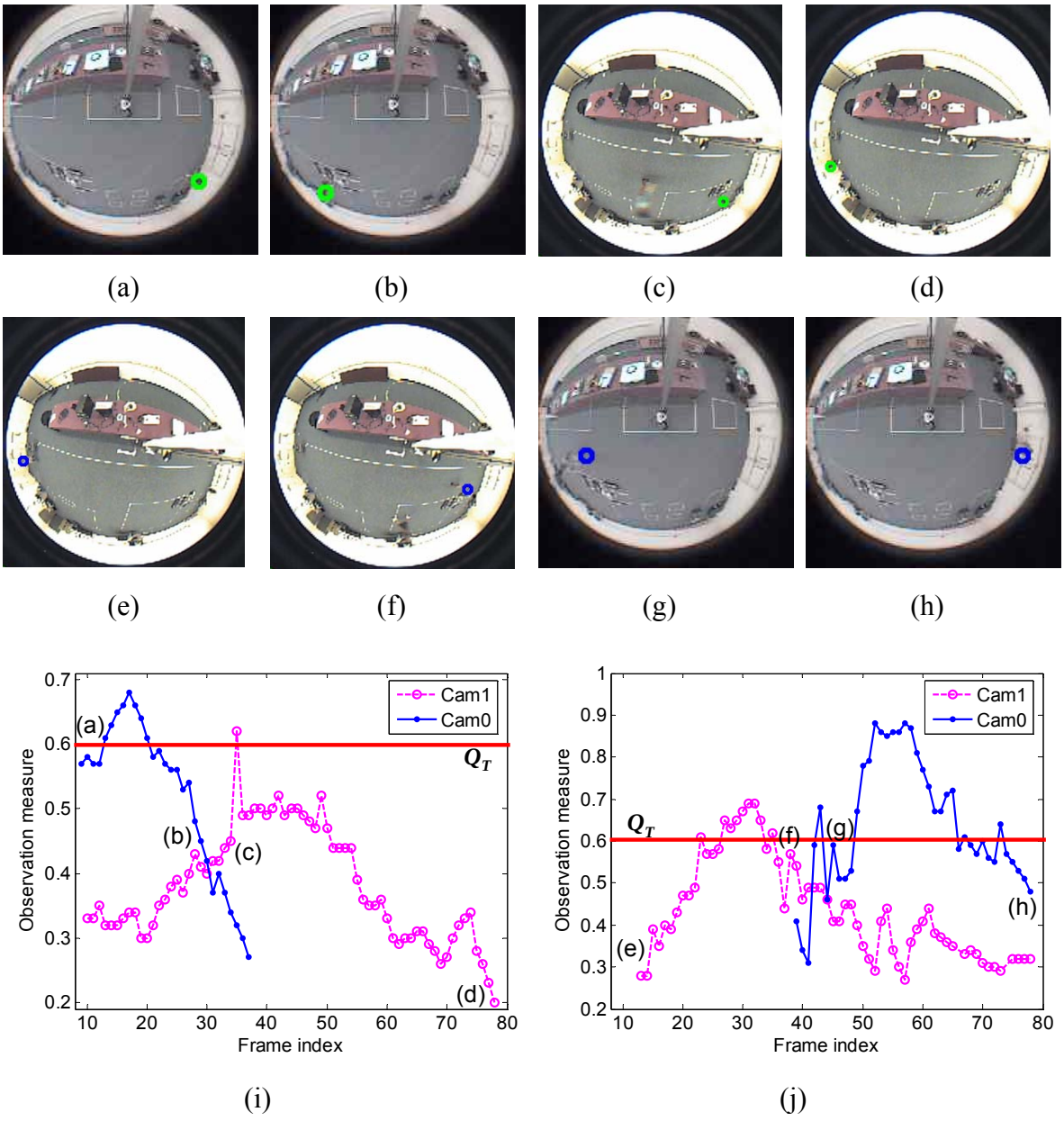
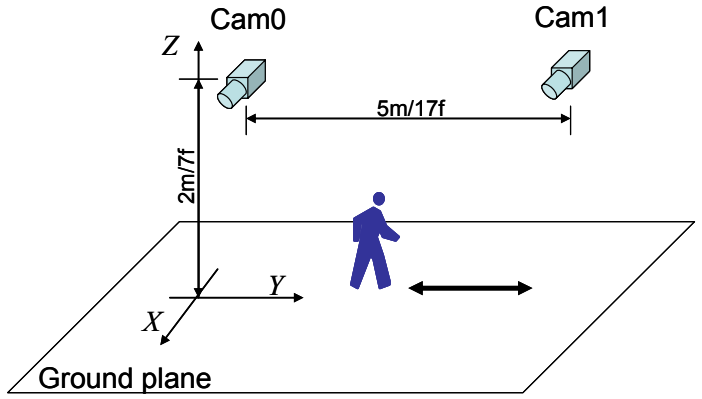
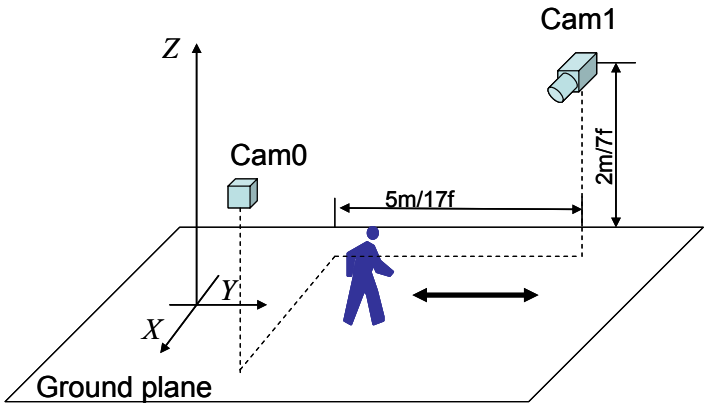


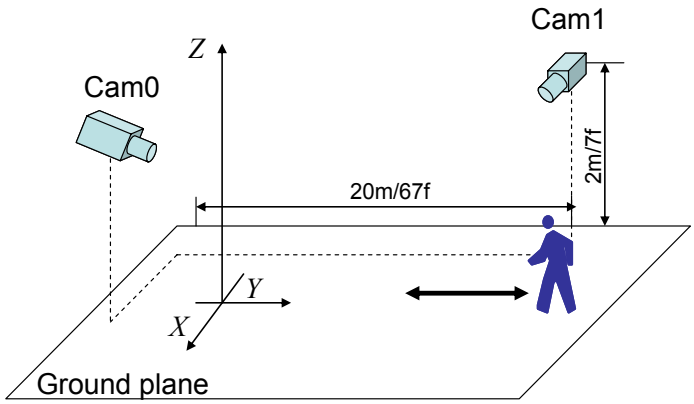
Figure 5.4. Camera handoff between omnidirectional cameras with dynamic occlusion: (a)-(d) target 0 and (e)-(h) target 1. (a) and (e) First frames with the detected targets. (b) and (f) Frames before camera handoff. (c) and (g) Frames after camera handoff. (d) and (h) Last frames before the targets become untraceable. (a), (b), (g), and (h) Frames captured by camera 0. (c), (d), (e), and (f) Frames captured by camera 1. Observation measure of (i) target 0 and (j) target 1.



(a)

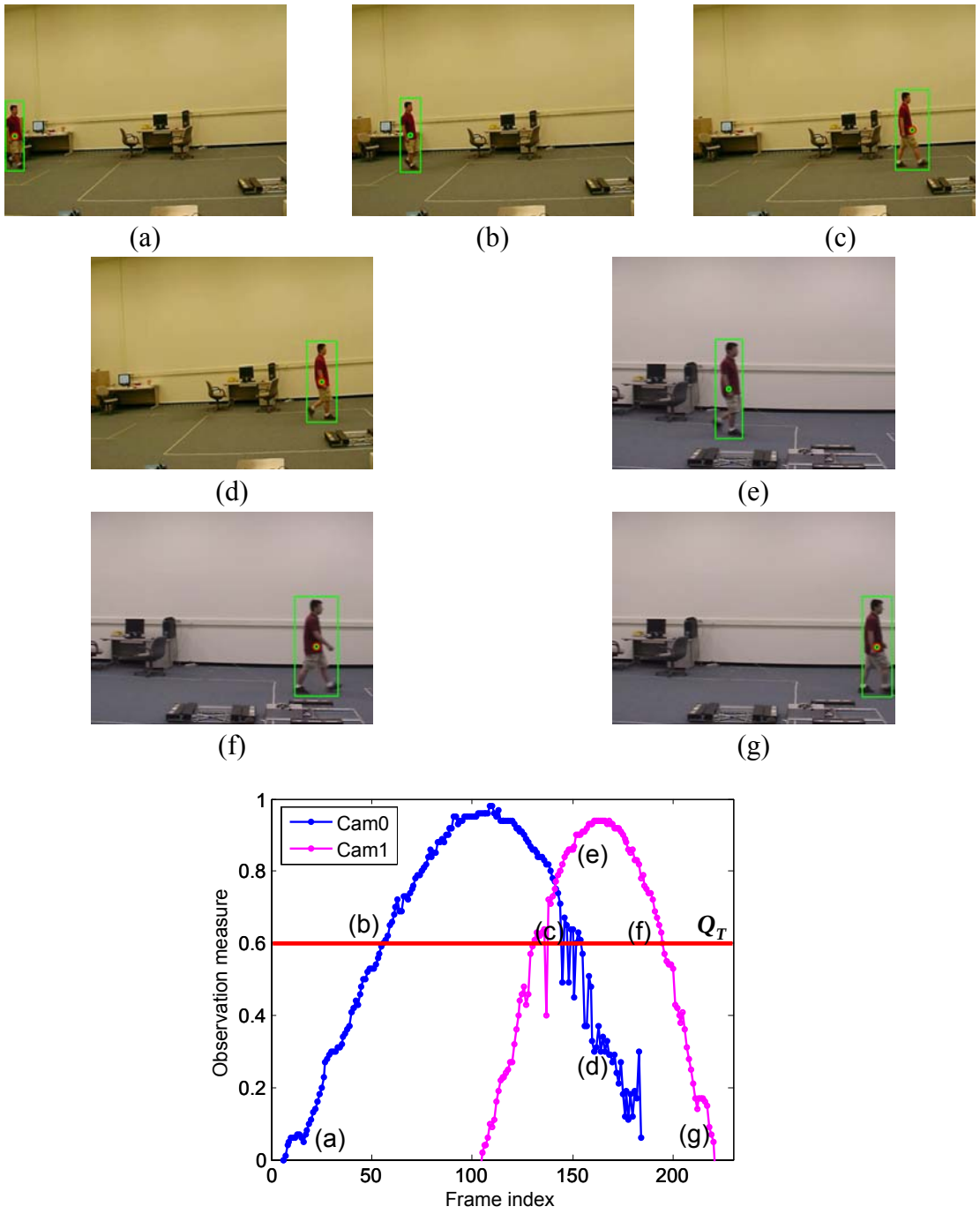


(b)



(c)

Figure 5.5. Schematic illustration of system setups for experiments using two static perspective cameras. Angles between the optical axes of the two cameras: (a) 0°, (b) 180°, and (c) 90°.



(h)

Figure 5.6. Camera handoff between static perspective cameras for case A. (a) First frame with the detected target in camera 0. (b) Tracked target with $Q \geq Q_T$ in camera 0. (c) Triggered handoff in camera 0. (d) and (e) Handoff is executed. The tracked target is transferred from camera 0 to camera 1. (f) Triggered handoff in camera 1. (g) Last frame before the target disappears from the camera's FOV. (h) Observation measure.

constant throughout the sequence. As expected, the observation measure increases as the target approaches the image center and decreases as the target moves away from the image center.

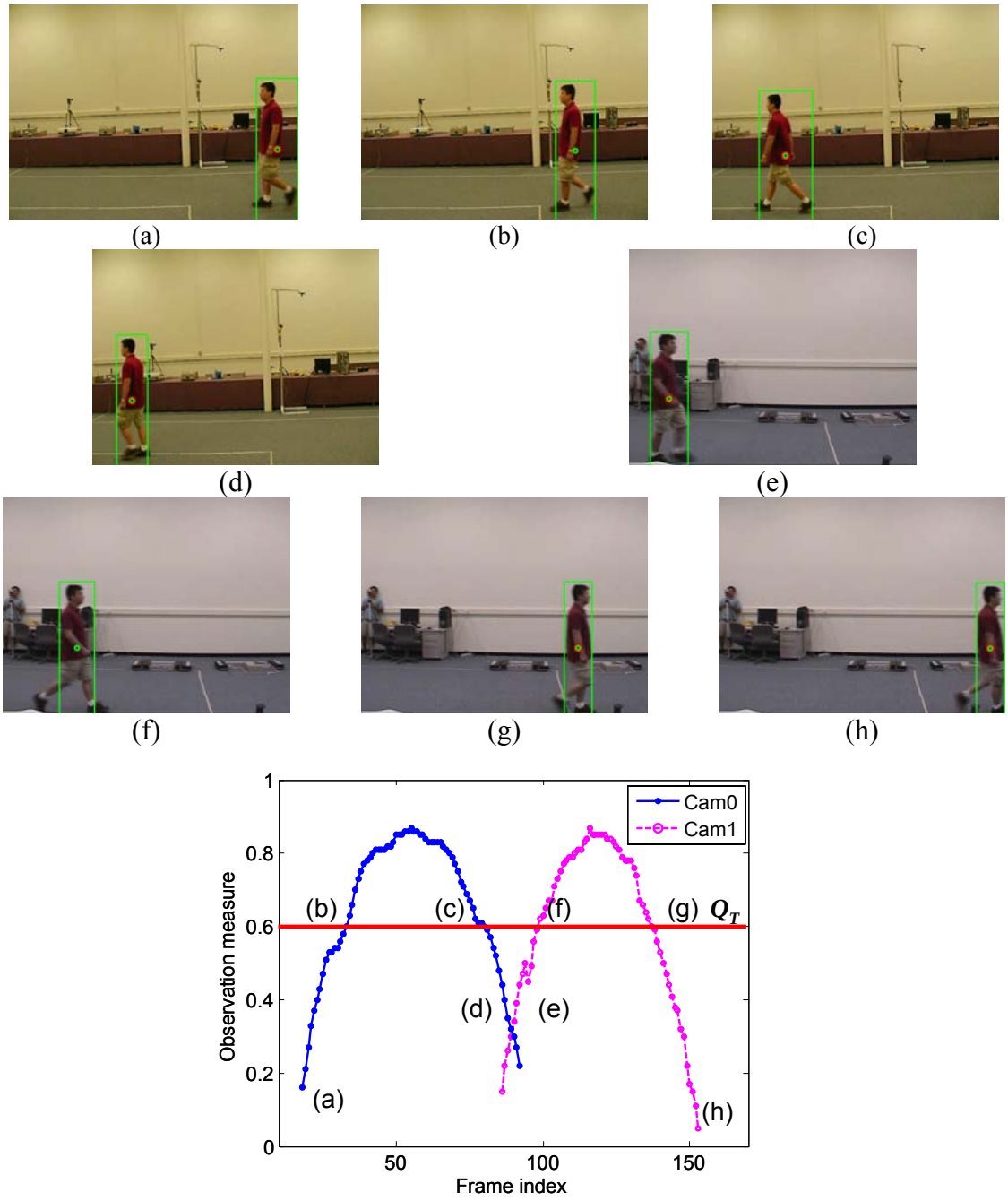
Similar experimental results are obtained for case B as shown in Figure 5.7, where the two cameras facing each other with an angle of 180° between their optical axes. The most significant factor in camera handoff is the M_D component, which triggers a handoff when the target is too close to the edges of the camera's FOV and chooses a camera with a target's image closer to the image center.

In case C, the resolution component M_R enters into effect as the target approaches camera 0 along the camera's optical axis. From Figure 5.8(h), we can see that the observation measure increases gradually as the target moves toward the camera. The target is transferred to camera 1 once it appears in the camera's FOV since a higher observation measure is achieved mainly because of the higher resolution in camera 1.

From our experiments, we could conclude that the proposed camera handoff algorithm is able to trigger a handoff request timely and select the suitable camera efficiently. The observation measure designed for camera handoff closely approximates the observation measure used for sensor planning so that sufficient time margins are reserved for communication, consistent labeling, and handoff execution. For all tested sequences, camera handoffs are carried out smoothly and successfully under scenarios with different system setups, varying resolution, color discrepancies, and partial occlusions.

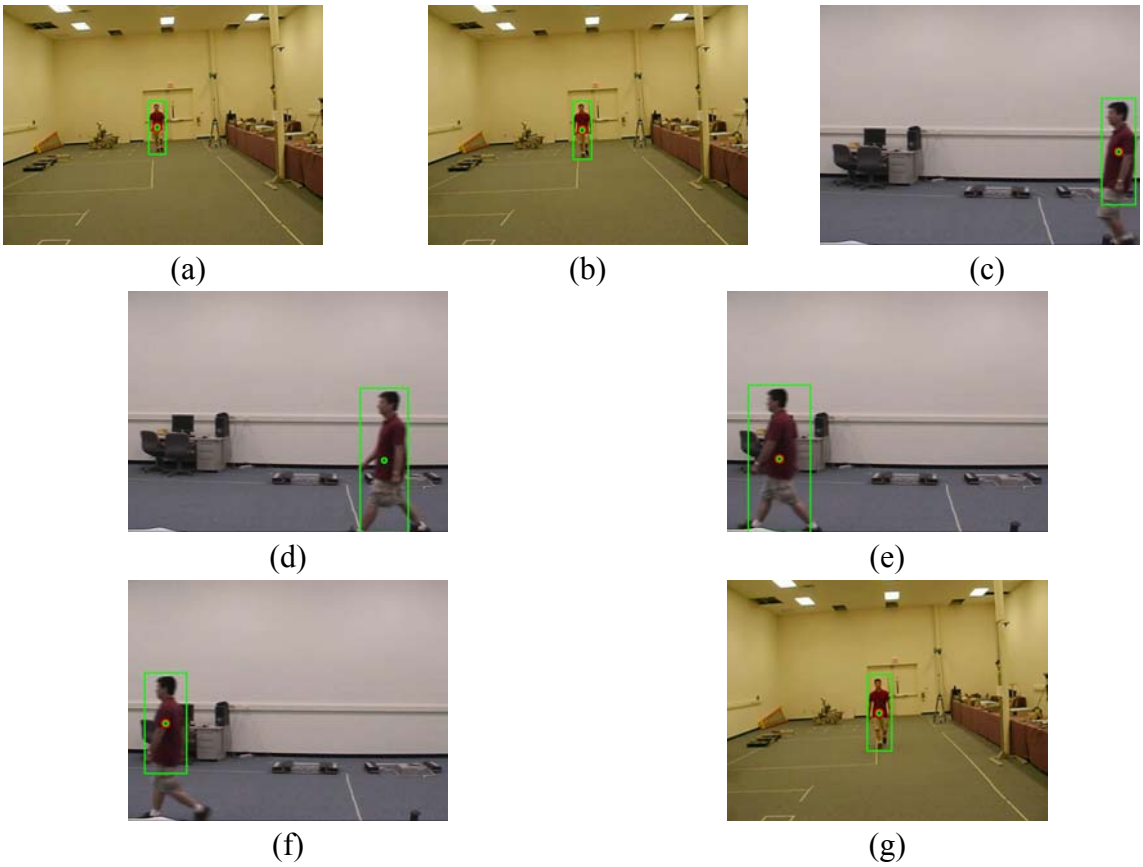
5.3.3 Camera handoff using synthetic data

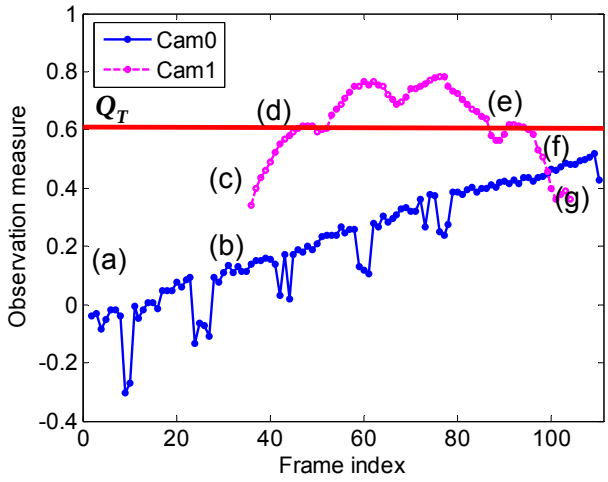
The aforementioned experiments are conducted using real-time pedestrian sequences, where the number of targets and the variety of sequences are limited due to the available experimental conditions. To obtain a statistically valid estimation of the handoff success rate, simulations are carried out to enable a large amount of tests under various conditions. In the following experiments, floor plan A with the camera placement optimized by the T1H method using static perspective cameras and floor plan B with the camera placement optimized by the T1P method using omnidirectional cameras are used. A pedestrian behavior simulator [Antonini06, Pettre02] is implemented with the arrival of the pedestrian following a Poisson distribution at an average arrival rate of 0.01 persons per second. The average walking speed of these generated objects is 0.5m per second. Several points of interest are generated randomly to form a pedestrian trace. The handoff success rate is obtained from simulation results of 300 randomly generated pedestrian traces. To verify the efficiency of the criterion ψ_{ij} defined in (5.12) in choosing a suitable camera in noisy applications, we manually add image noise. The noise standard deviation is varied from 5 to 25 pixels. System performance is compared based on the handoff success rate using the noisy observation measure Q_{ij} and the modified quantity ψ_{ij} as the criterion for camera handoff.



(i)

Figure 5.7. Camera handoff between static perspective cameras for case B. (a) First frame with the detected target in camera 0. (b) Tracked target with $Q \geq Q_T$ in camera 0. (c) Triggered handoff in camera 0. (d) and (e) Handoff is executed. The tracked target is transferred from camera 0 to camera 1. (f) Tracked target with $Q \geq Q_T$ in camera 1. (g) Triggered handoff in camera 1. (h) Last frame before the target disappears from the camera's FOV. (i) Observation measure.





(h)

Figure 5.8. Camera handoff between static perspective cameras for case C. (a) First frame with the detected target in camera 0. (b) and (c) Handoff is executed. The tracked target is transferred from camera 0 to camera 1. (d) Tracked target with $Q \geq Q_T$ in camera 1. (e) Triggered handoff in camera 1. (f) and (g) Handoff is executed. The tracked target is transferred from camera 1 to camera 0. (h) Observation measure.

Figure 5.9 shows the optimal camera placement and the handoff success rate with Q_{ij} and ψ_{ij} as the criterion for camera transition. For static perspective cameras as shown in Figure 5.9(b), the handoff success rate is maintained disregarding the significantly increased noise level when ψ_{ij} is used. In comparison, the handoff success rate drops gradually from 84.6% to 51.3% when Q_{ij} is used. Similar observations apply to the experiments based on omnidirectional cameras as shown in Figure 5.9(d). The handoff success rate is maintained for ψ_{ij} while the handoff success rate degrades from 98.5% to 86.4% for Q_{ij}. Therefore, in practical surveillance, the modified quantity ψ_{ij} is a robust criterion for camera selection. An approximately constant handoff success rate is achieved disregarding the system's noise level.

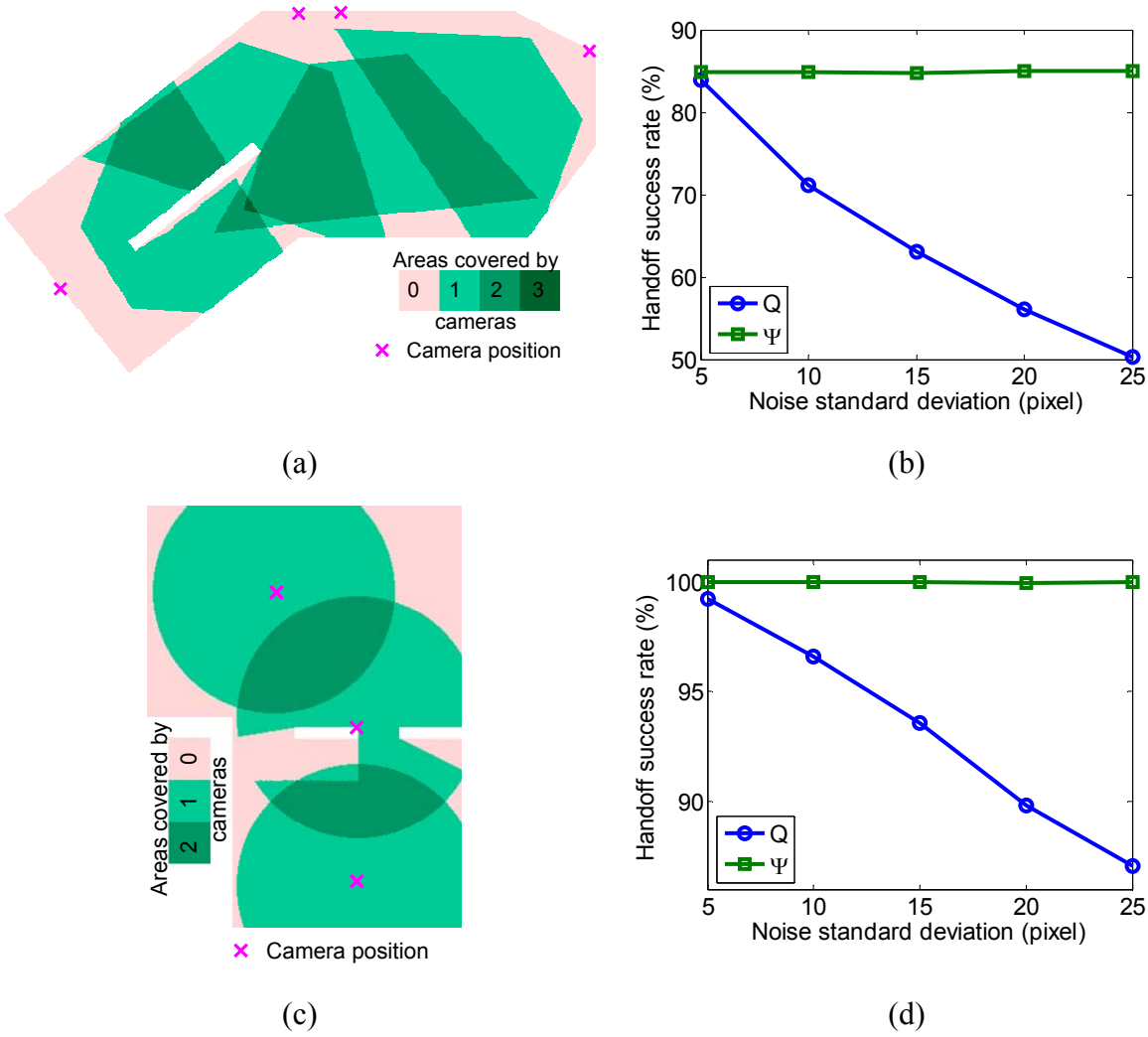


Figure 5.9. (a) The optimal camera placement obtained based on the T1H method for floor plan A using static perspective cameras. (b) Comparison of the handoff success rate based on the camera placement in (a) when the noisy observation measure Q_{ij} and the modified quantity ψ_{ij} are used. (c) The optimal camera placement obtained based on the T1P method for floor plan B using omnidirectional cameras. (d) Comparison of the handoff success rate based on the camera placement in (c) when the noisy observation measure Q_{ij} and the modified quantity ψ_{ij} are used.

6 High magnification face recognition

Long observation distances and high optical magnifications introduce severe and nonuniform image blur. To quantify the degree of magnification blur, we look into conventional sharpness measures, which are widely used to evaluate out-of-focus blur. Our study shows that conventional sharpness measures are sensitive to image noise and therefore are not suitable for our applications. A class of adaptive sharpness measures is proposed to suppress artificially elevated sharpness values due to image noise by assigning nonlinear weights to the image gradients according to their local activities. After assessing magnification blur, we investigate the degradations in FRR introduced by magnification blur and verify that magnification blur is another major degrading factor in addition to pose and illumination. Image enhancement is a common practice to improve image quality. In this chapter, we implement several backbone enhancement algorithms and compare their performances based on the FRR of the processed face images. A wavelet based algorithm is chosen for the restoration of high magnification face images because of its ability to balance noise reduction and detail enhancement.

The remainder of this chapter is organized as follows. Section 6.1 introduces our adaptive sharpness measures with experimental validation. The long range high magnification face database (UT-LRHM) is described in section 6.2. Section 6.3 presents our face image quality assessment method. The wavelet based enhancement algorithms are studied and a significantly improved FRR is demonstrated in section 6.4.

6.1 Adaptive sharpness measure

Sharpness measures have been traditionally proposed to evaluate out-of-focus blur. However, conventional sharpness measures are sensitive to image noise. Since the image noise level increases as the system magnification increases, conventional sharpness measures are not directly applicable to high magnification images. To avoid artificially elevated sharpness values due to image noise, adaptive measures are proposed [Yao06A]. In order to differentiate between variations caused by actual image edges and those introduced by image noise and artifacts, adaptive sharpness measures assign different weights to pixel gradients according to their local activities. For pixels in smooth areas, small weights are used. For pixels adjacent to strong edges, large weights are allocated.

6.1.1 Definition

The definition of local activities and the selection of weight functions are two major factors in constructing adaptive sharpness measures. According to the description of local activities, sharpness measures can be divided into two groups: separable and non-separable. As the name suggests, separable methods consider horizontal and vertical edges independently while non-separable methods include the contributions from diagonal edges. For separable measures, two weight signals are constructed, a vertical

$$g_x(x,y) = f(x+1,y) - f(x-1,y)\,, \tag{6.1}$$

and a horizontal

$$g_y(x,y) = f(x,y+1) - f(x,y-1)\,. \tag{6.2}$$

For non-separable methods, the weight signals are:

$$g(x,y) = f(x-1,y) + f(x+1,y) - f(x,y-1) - f(x,y+1)\,. \tag{6.3}$$

Different forms of weight functions can be used, among which polynomial and rational functions are two popular choices. Polynomial and rational functions are also exploited in adaptive unsharp masking [Ramponi98A, Ramponi98B]. The polynomial weights suppress small variations mostly introduced by image noise and have proved efficient in evaluating the sharpness of high magnification images [Yao06A]. The rational weights emphasize a particular range of image gradients. Considering the non-separable method $g(x,y)$ for example, the polynomial weight function is given by:

$$\omega(x,y) = g(x,y)^{p_\omega}\,, \tag{6.4}$$

where p_ω is the power index determining the degree of noise suppression. The rational weight function can be written as:

$$\omega(x,y) = \frac{(2k_0 + k_1)g(x,y)}{g^2(x,y) + k_1 g(x,y) + k_0^2}\,, \tag{6.5}$$

where k_0 and k_1 are coefficients associated with the peak position L_0 and width ΔL of the response, respectively, and comply with the following relation $k_0 = L_0$ and $k_1^2 + 8k_0k_1 + 12k_0^2 - \Delta L^2 = 0$. Figure 6.1 illustrates the comparison of different forms of weights.

These weights are then applied to gradient based sharpness measures to construct adaptive sharpness measures. Considering the Tenengrad sharpness measure [Krotkov89] for instance, the corresponding separable measure is given by:

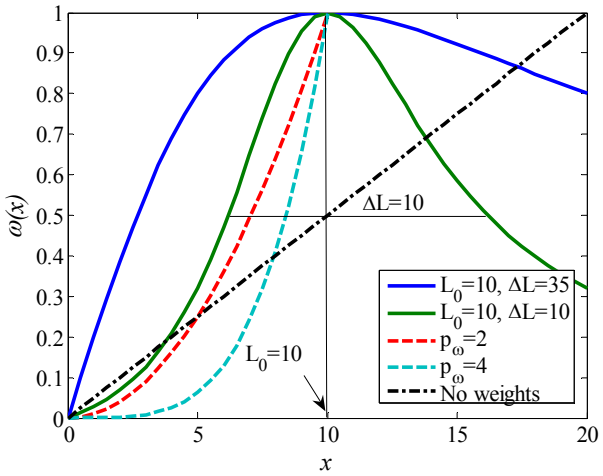


Figure 6.1. Illustration of weight functions. Solid curves: rational functions and dashed curved: polynomial functions.

$$S = \sum_{x=1}^{N_{row}} \sum_{y=1}^{N_{col}} \left[\omega_x(x,y) f_x^2(x,y) + \omega_y(x,y) f_y^2(x,y) \right], \tag{6.6}$$

where $\omega_x(x,y) / \omega_y(x,y)$ denotes the weights obtained from $g_x(x,y) / g_y(x,y)$, N_{row}/N_{col} is the number of image rows/columns, and $f_x(x,y) / f_y(x,y)$ represents the vertical / horizontal gradient at pixel (x,y) obtained via the Sobel filter. For non-separable methods, the adaptive Tenengrad is formulated as:

$$S = \sum_{x=1}^{N_{row}} \sum_{y=1}^{N_{col}} \omega(x,y) \left[f_x^2(x,y) + f_y^2(x,y) \right]. \tag{6.7}$$

The newly developed adaptive sharpness measures assign different weights to each pixel according to its local activity. This modification avoids measuring noise, enhances the responses from image edges, and thus results in a robust performance in noisy applications. Moreover, since no edge detection and parameterization are involved, the computational cost remains low.

6.1.2 Experimental results

We first validate the definition of the adaptive sharpness measures by examining their responses to out-of-focus blur. Three sequences, referred to as resolution chart (RC), license plate (LP), and man's face (MF), are collected using a Canon A80 camera at intervals of three focus motor steps covering a focus range of 0.2m to infinity (a total of 53 images per sequence). Other camera configurations, such as zoom, iris, shutter speed,

and exposure compensation, are kept unchanged. Figure 6.2 shows sample images collected at the best focus.

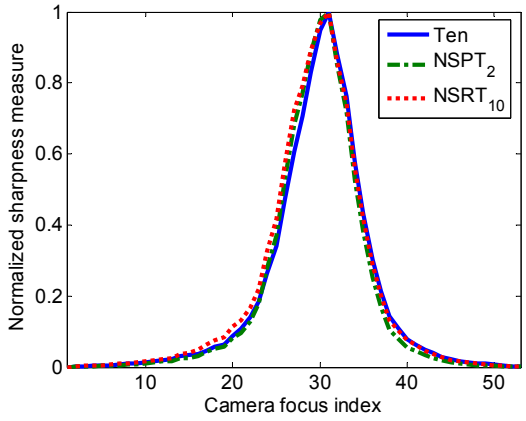
We implement two types of adaptive sharpness measures: non-separable polynomial Tenengrad with $p_\omega = 2$ ($NSPT_2$) and non-separable rational Tenengrad with $L_0 = 10$ and $\Delta L = 10$ ($NSRT_{10}$). The performances of these two measures are studied with respect to a varying camera focus in Figure 6.3. For the RC and LP sequences shown in Figures 6.3(a) and (b), the performances of the Tenengrad and the adaptive Tenengrad are similar. Performance differences are observed in the MF sequence. From the $NSPT_2$ measure, two local maxima are obtained corresponding to the optimal focus positions of the foreground (face) and the background (brick wall). The conventional Tenengrad measure only captures the focus position of the background, since the brick wall contains stronger and denser edges. The $NSRT_{10}$ measure is unable to detect two focus planes either. However, this can be corrected by choosing a different set of L_0 and ΔL.

To verify the improved robustness to image noise, Gaussian noise with a standard deviation varying from 1 to 20 grey levels is added to the original images. Figure 6.4 summarizes the performances of the conventional Tenengrad and the adaptive Tenengrad measures. The response of the conventional Tenengrad maintains the desired shape at all noise levels. However, for a given focus, its value increases as the noise level increases, resulting in a set of shifted curves. It is evident that the Tenengrad measure is unable to differentiate variations induced by noise from those introduced by the actual changes in focus. Taking the MF sequence in Figure 6.4(g) as an example, the sharpness value with a focus index of 35 and a noise standard deviation of 10 is the same as the one with a focus index of 45 and a noise standard deviation of 5. These spurious variations are the result of noise and should be eliminated. In comparison, as the camera's focus approaches its optimal position, the adaptive Tenengrad measures begin to respond in a different manner and their values decrease with respect to increased noise level. The adaptive Tenengrad measures can counteract the fluctuations caused by noise and respond in a manner agreeable with visual perception.

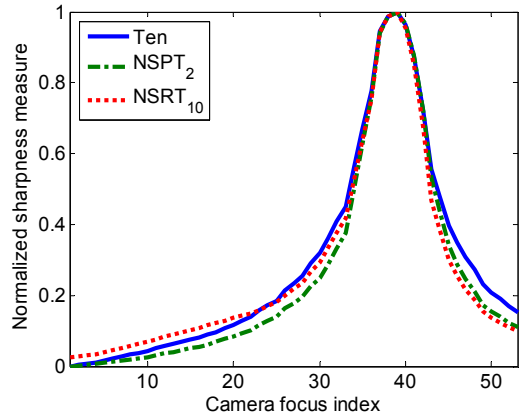
Thresholding can be used in the conventional Tenengrad to reduce the influence of image noise. However, since more and more pixels are regarded as noisy and are eliminated from successive processing as image noise increases, the accuracy of the resulting sharpness measure suffers considerably. The thresholded Tenengrad sharpness

(a) (b) (c)

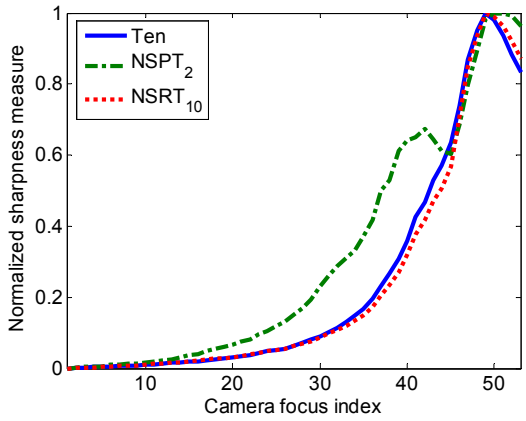
Figure 6.2. Sample frames from the tested sequences: (a) resolution chart, (b) license plate, and (c) men's face.



(a)



(b)



(c)

Figure 6.3. The performance of the Tenengrad (Ten) and adaptive Tenengrad sharpness measures ($NSPT_2$ and $NSRT_{10}$) with respect to a varying camera focus. The focus index represents samples of the camera's focus at intervals of three motor steps. (a) Resolution chart, (b) license plate, and (c) man's face.

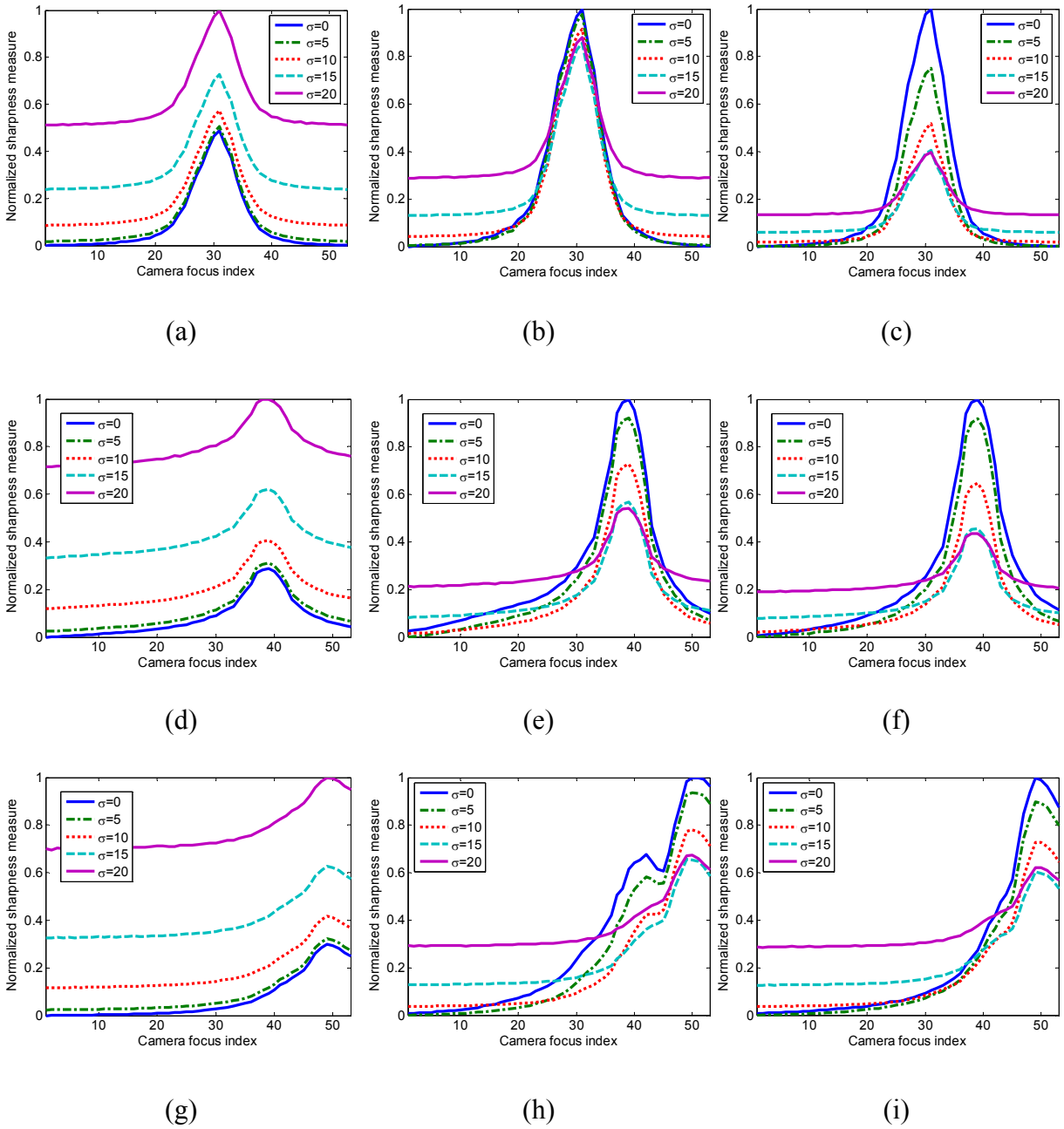


Figure 6.4. Sharpness measures of images corrupted by additive Gaussian noise. (a)-(c): Resolution chart. (d)-(f): License plate. (g)-(i): Men's face. (a), (d), and (g): Conventional Tenengrad. (b), (e), and (h): $NSPT_2$. (c), (f), and (i): $NSRT_{10}$. σ denotes the standard deviation of noise.

measure sacrifices accuracy to neutralize variations caused by noise. This loss of information is frequently substantial to where the accuracy of the resulting sharpness measure deteriorates. Therefore, even with proper thresholding, the Tenengrad measure is unable to achieve comparable robustness to noise as the newly designed adaptive Tenengrad measure. Furthermore, the selection of the threshold depends on the image noise level and if the proper threshold is to be obtained, the image noise level should be estimated first. A small threshold is unable to balance the noise, while a large threshold results in a considerable information loss. In comparison, the adaptive Tenengrad is able to automatically adjust the weights for every pixel without prior knowledge of the image noise.

6.2 Face database

Our database collection, including indoor and outdoor sessions, began in February 2006 and ended in October 2006. The data set contains frontal view face images and videos collected with various system magnifications (10×~284×), observation distances (10m~300m), indoor (office ceiling light and side light) and outdoor (sunny and cloudy) illuminations, still/moving subjects, and constant/varying camera zooms. Small expression and pose variations are also included in the video sequences of our database, as shown in Figure 6.5, closely resembling the variations encountered in uncontrolled surveillance applications.

For the indoor sequence collection, the observation distance is varied from 10m to 16m. Given this distance range and an image resolution of 640×480, a 22× optical magnification is sufficient to yield a face image with an inter-ocular distance of 60 pixels. This resolution is recommended by FaceIt® for successful recognition. Therefore, a commercially available PTZ camera (Panasonic WV-CS854) was used.

Our indoor database includes both still images (eight images per subject) and video sequences (six sequences per subject). Still images are collected at uniformly distributed distances in the range of 10m to 16m with an increment of 1m approximately. The corresponding system magnification varies from 10× to 20× with an increment of 2×,

(a) (b)

Figure 6.5. Illustration of (a) a small expression change and (b) a small pose variation.

achieving an approximately constant inter-ocular distance to eliminate the effect of resolution. Still images with low magnification (1×) are also taken from a close distance (1m) as a reference image set. The achievable face recognition rate using this image set provides an ideal performance reference for evaluating degradations caused by high magnification.

The observation distance and system magnification are two major factors, to which this effort is devoted. Meanwhile, the effect of composite target and camera motions is included to achieve a close resemblance to practical surveillance scenarios. Therefore, the indoor video sequences are recorded under the following conditions: (1) constant distance & varying system magnification, (2) varying distance (the subject walks at a normal speed toward the observation camera) & constant system magnification, and (3) varying distance & varying system magnification. Conditions 1 and 2 concentrate on the individual effect of camera zoom and subject motion, respectively, while the combined effect can be observed in condition 3. In addition, the system magnification in condition 3 is varied so that a constant inter-ocular distance is maintained. These video sequences can be used for the studies regarding the effect of resolution, subject motion, and camera zoom. Figure 6.6 shows example face images degraded by blurs from the subject's motion, the camera's zoom motion, and the camera's focus motion.

The aforementioned still images and video sequences are collected under fluorescent ceiling lights with full intensity (approximately 500Lux) and include a certain degree of

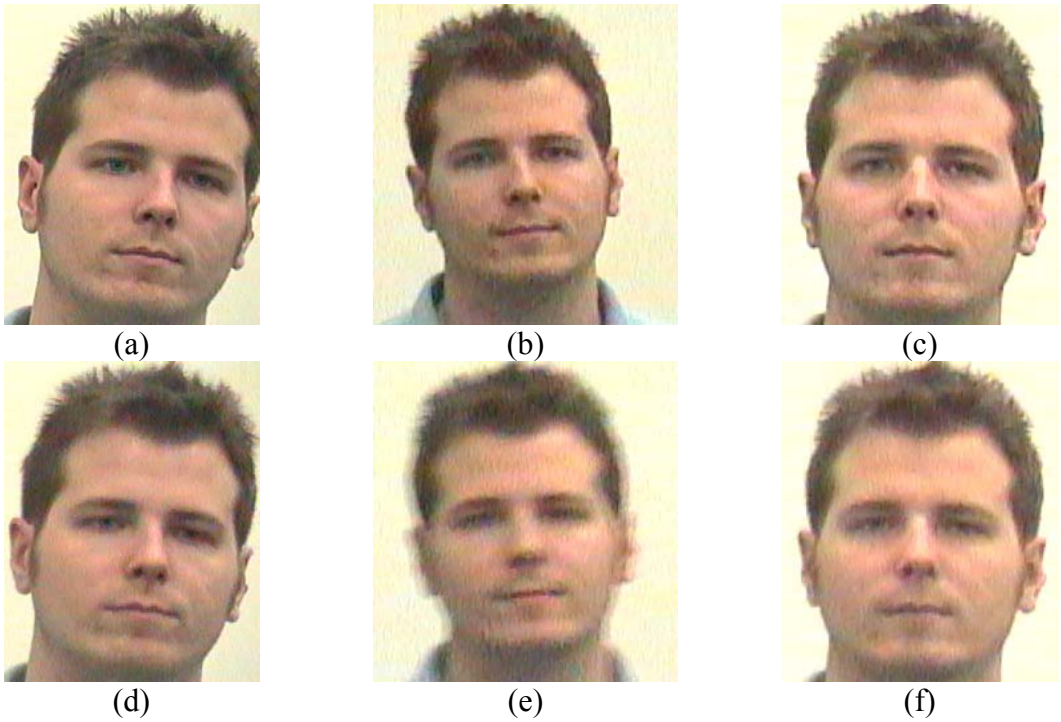


Figure 6.6. Illustration of various types of blurs captured in our database in addition to magnification blur. (a)-(c) Reference images. Blurred images due to: (d) subject's motion, (e) camera's zoom motion, and (f) camera's focus motion.

illumination changes caused by a varied distribution of the ceiling lights. Our indoor database also considers a large amount of illumination changes. A halogen side light (approximately 2500Lux) is added and a sequence is recorded as the intensity of the ceiling lights is decreased from 100% to zero, which creates a visual effect of a rotating light source.

The gallery images are collected by a Canon A80 camera under a controlled indoor environment from a distance of 0.5m. The image resolution is 2272×1704 pixels and the camera's focal length is 114mm (magnification: 2.28×). Figures 6.7 and 6.8 illustrate sample images of one data record in the database. A data record is a series of images of a given subject under all shooting conditions. Table 6.1 summarizes the specifications of the indoor data sets.

The indoor session has 55 participants (78% male and 22% female). Their ethnic distribution consists of 73% Caucasian, 13% Asian, 9% Asian Indian, and 5% of African descent. The image resolution is 640×480 pixels. For the video sequences, our database provides uncompressed frames in the format of BMP files at a rate of 30 frames per second as well as AVI files compressed using Microsoft MPEG 2.0 codec. Each video sequence lasts 9 seconds. The total physical size for storage is 84 GB, with 1.53 GB per subject.

For the outdoor sequence collection, a composite imaging system was built where a Meade ETX-90 telescope (focal length: 1250mm) was coupled with a JVC MG-37U camcorder (focal length: 2.3-73.6mm) via a Celestron 40mm eyepiece using an afocal connection. The achievable system magnification is of 22×~659×.

Our outdoor database includes both still images (two images per subject) and video sequences (twelve sequences per subject). Two sequences per subject are collected at uniformly distributed distances from 50m to 300m with an increment of 50m. The corresponding system magnification varies from 66× to 284× with an increment of about 44×, achieving an approximately constant inter-ocular distance. The two sequences are collected with different subject motions, one with the subject standing still and the other with the subject walking a short distance. One still image per subject is also collected from a close distance (1m at 1×) for a reference image set. The gallery images are collected by a Nikon camera under a controlled indoor environment from a distance of 1m. The image resolution is 2560x1920 pixels. Figure 6.9 illustrates sample images of one data record in the outdoor database and Table 6.2 summarizes the specifications of the data sets.

6.3 Face image quality assessment

The following experiments are carried out using the UT-LRHM database described in section 6.2. The noise characteristics of the face images at various magnifications are studied. The standard deviation of a uniform background patch closely describes the noise behavior and is computed with respect to system magnification, as shown in Figure 6.10. The image noise increases as the system magnification changes from 1× to 20×.

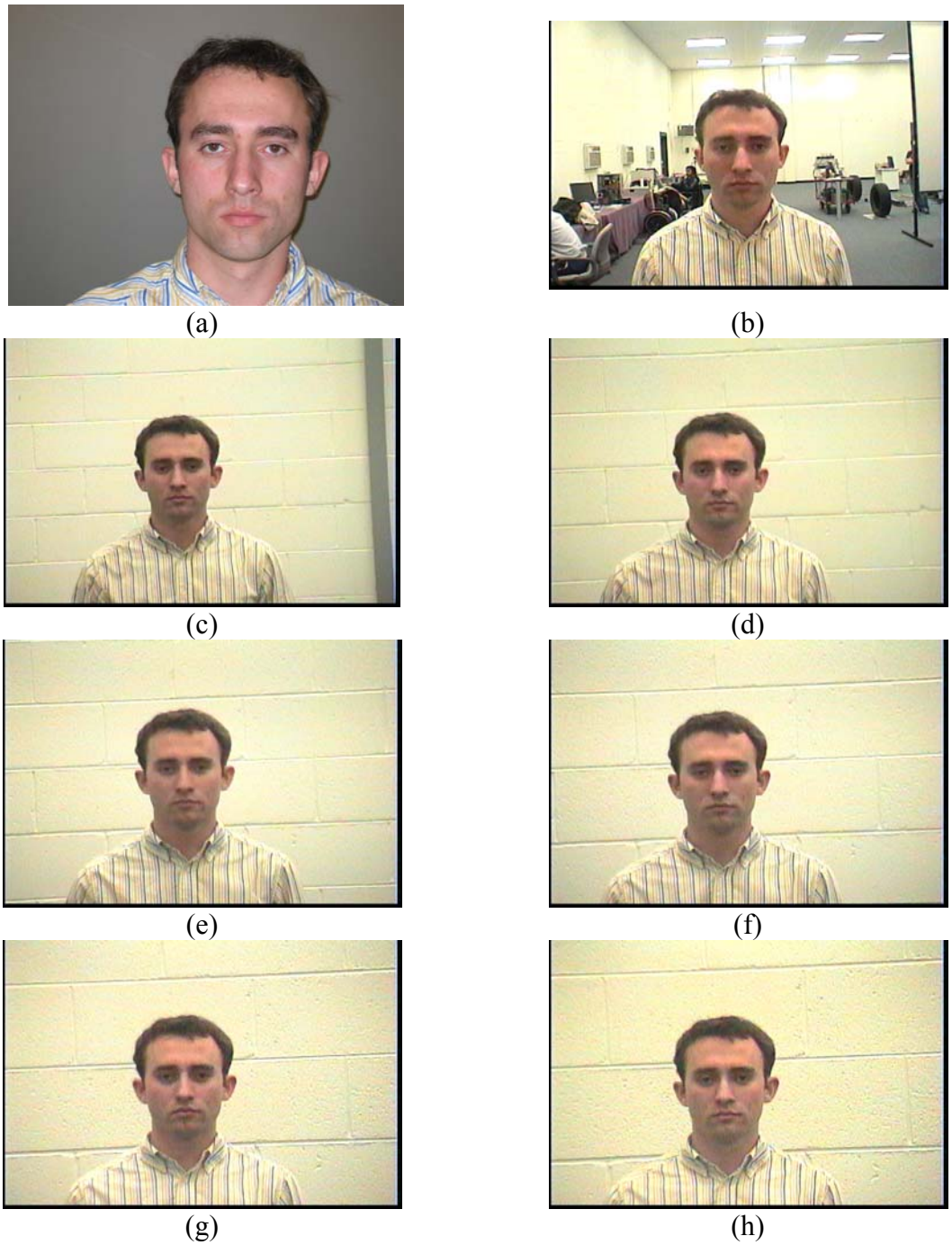


Figure 6.7. A set of still images in one data record from the indoor database: (a) gallery image, (b) 1× reference, 1m, 60p, (c) 10× , 9.5m, 57p, (d) 12×, 10.4m, 57p, (e) 14×, 11.9m, 58p, (f) 16×, 13.4m, 60p, (g) 18×, 14.6m, 60p, and (h) 20×, 15.9m, 60p. Face images in (b)-(h) have approximately the same resolution with an inter-ocular distance around 60 pixels. The inter-ocular distance is obtained by averaging all the face images across different subjects in each data set.

Figure 6.8. A set of sample frames from the collected sequences in one data record from the indoor database. (a) Condition 1: 20×→10×, 13.4m, constant observation distance. (b) Condition 2: 10×, 15.9m→9.5m, constant system magnification. (c) Condition 3: 20×→10×, 15.9m→9.5m, constant inter-ocular distance. (d) Varying illumination, 20×, 15.9m.

Figure 6.9. A set of sample frames from the standing sequences in one data record from the outdoor database: (a) indoor gallery image, (b) 1× reference, (c) 66× , 50m, 79p, (d) 109×, 100m, 76p, (e) 153×, 150m, 79p, (f) 197×, 200m, 76p, (g) 241×, 250m, 78p, and (h) 284×, 300m, 78p. Face images in (c)-(h) have approximately the same resolution with an inter-ocular distance of 80 pixels.

Table 6.1. The specifications of the indoor data sets.

Still images							
Magnification (×)	1	10	12	14	16	18	20
Distance (m)	1	9.5	10.4	11.9	13.4	14.6	15.9
Inter-ocular distance (pixel)	60	57	57	58	60	60	60

Video sequences		
Conditions	Mag. (×)	Distance (m)
1. Constant distance & varying system Mag.	20→10	13.4 and 15.9
2. Varying distance & constant system Mag.	10 and 15	15.9→9.5
3. Varying distance & varying system Mag.	20→10	15.9→9.5
Varying illumination, constant distance & system Mag.	20	15.9

Table 6.2. The specifications of the outdoor data sets.

Magnification (×)	66	109	153	197	241	284
Distance (m)	50	100	150	200	250	300
Inter-ocular distance (pixel)	79	76	79	76	78	78

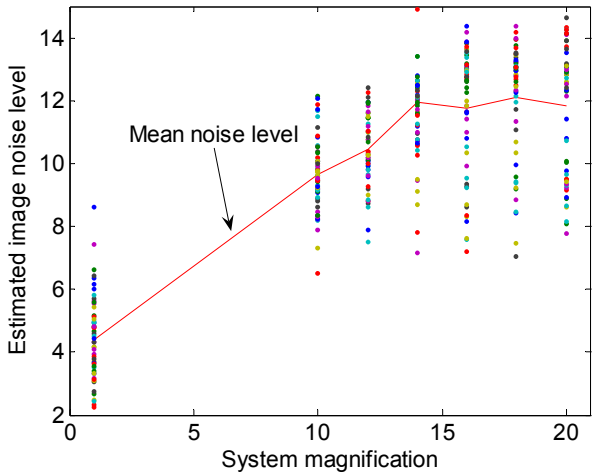


Figure 6.10. Face image noise level vs. system magnification. Image gray level: 0-255. Dots represent the standard deviation of the noise computed from face images of different subjects. The mean noise level increases as the system magnification increases.

Therefore, to exclude an artificially elevated sharpness value from increased image noise, adaptive sharpness measures are used.

In the following experiments, each data set consists of face images collected from the same observation distance and with the same system magnification. The sharpness measures of these face images are computed and the mean sharpness values are obtained by averaging them across different subjects within one data set. Figure 6.11 shows the computed sharpness values ($NSPT_2$) and their means. These mean sharpness values present a clearer view of how image quality responds to system magnification and observation distance. As expected, image sharpness decreases gradually as the system magnification and observation distance increase for both indoor and outdoor image sets.

Now we study the influence of magnification blur on face recognition rate. The gallery image sets are compared against different sets of probe images with an approximately constant inter-ocular distance of 60 pixels, each set consisting of face images collected at the same observation distance and with the same system magnification. The face recognition rate at various system magnifications is illustrated in Figure 6.12 and Table 6.3. It is obvious that image deterioration from limited fine facial details causes the FRR to drop gradually as the system magnification increases. For the indoor session, the CMCM declines from 69.7% to 58.8% as the system magnification increases from 10× to 20×. There exists a significant performance gap between the probes with low (1×) and high (20×) magnifications. Similar observations apply to the outdoor session, where the CMCM declines from 64.5% to 42.6% as the observation distance increases from 50m to 300m. This reveals that magnification blur is a major degrading factor in face recognition and that the performance gap between image sets with low and high magnifications is to be compensated for by image enhancement.

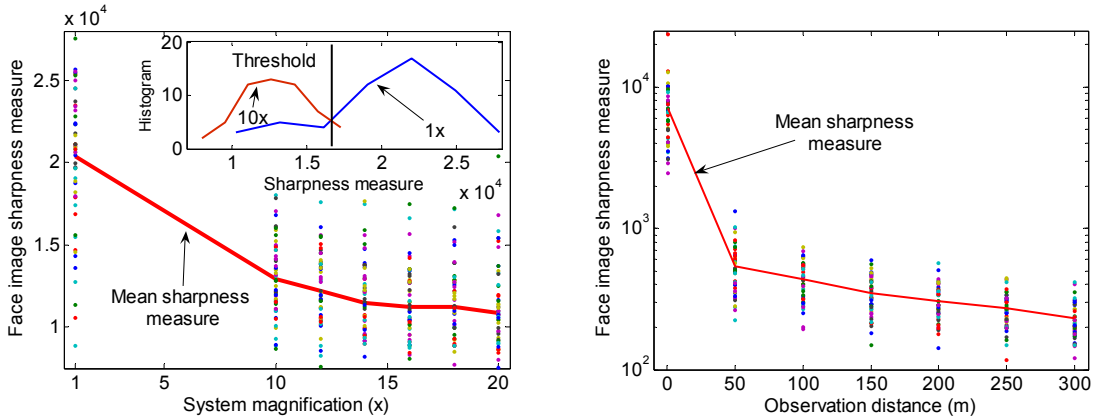
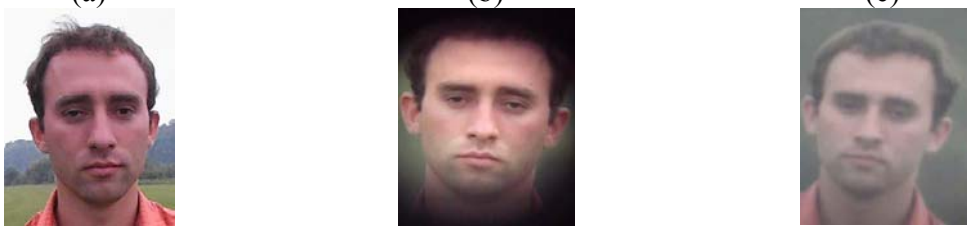
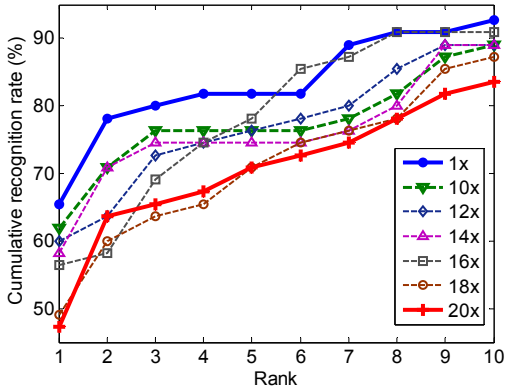


Figure 6.11. Sharpness measures for face images collected with different system magnifications/observation distances: (a) indoor and (b) outdoor sessions. Dots represent the sharpness measures computed from face images of different subjects. The mean sharpness measure decreases as the system magnification/observation distance increases.

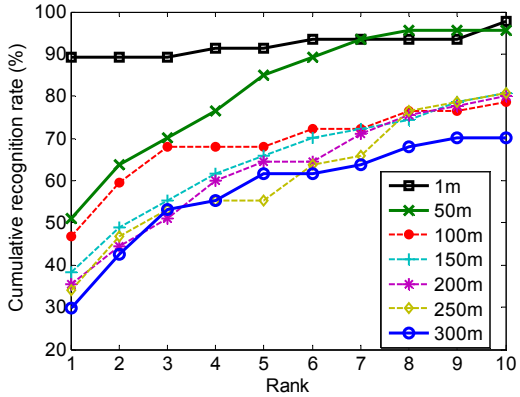
(a) (b) (c)



(d) (e) (f)



(g)



(h)

Figure 6.12. Sample images from the indoor session at different system magnifications: (a) 1×, (b) 10×, and (c) 20×. Sample images from the outdoor session at different observation distances: (d) 1m, (e) 50m, and (f) 300m. CMC comparison across probe sets with different system magnifications and observation distances: (g) indoor and (h) outdoor sessions. FRR drops gradually as the system magnification / observation distance increases.

Table 6.3. Performance comparison based on CMCM and rank-one recognition rate across system magnifications and observation distances.

Indoor session							
System magnification (×)	**1**	**10**	12	14	16	18	**20**
CMCM (%)	**74.3**	**69.7**	67.3	67.5	64.9	59.4	**58.8**
Rank-one (%)	**65.5**	**61.8**	60.0	58.2	56.4	49.1	**47.3**
Outdoor session							
Observation distance (m)	**1m**	**50m**	100m	150m	200m	250m	**300m**
CMCM (%)	**90.7**	**64.5**	57.5	50.2	47.3	46.6	**42.6**
Rank-one (%)	**89.4**	**51.1**	46.8	38.3	35.6	34.0	**29.8**

The decrease in FRR caused by magnification blur is consistent with the behavior of image sharpness measures shown in Figure 6.11. Therefore, we could use the adaptive Tenengrad as an indicator not only for the degree of magnification blur but also for recognition rate. From the distribution of these sharpness values, especially those of the 1× and 10× image sets, a threshold (the intersection point in Figure 6.11) can be derived, $S_{th} = 16600$, which separates the tested face images into two groups: one with acceptable sharpness ($S \geq S_{th}$) and the other degraded by magnification blur ($S < S_{th}$). Images in the first group contain sufficient facial features and thus will not deteriorate the overall FRR. On the contrary, images in the second group, deficient in necessary facial features, require image enhancement so that the overall FRR can be maintained. The threshold $S_{th} = 16600$ is obtained empirically and is application dependent. In practice, the sharpness measures of low magnification images can be computed and their statistics, such as the mean S_0 and the standard deviation σ_s, can be estimated. The threshold can then be defined as $S_{th} = S_0 - \sigma_s$. The threshold can also be estimated and updated on-the-fly by studying the distributions of image sharpness at various magnifications.

6.4 Enhancement of high magnification face images

As illustrated in section 6.3, high magnification images suffer from increased image noise and magnification blur. In general, deblurring algorithms increase image noise, while denoising algorithms smooth out image details. The resulting images are either short of details or overwhelmed by elevated image noise. Since FaceIt® is sensitive to both types of degradation, a good balance is to be found for an optimal FRR. Multi-scale processing based on wavelet transform is used and proves effective. After wavelet decomposition, the vertical, horizontal, and diagonal detail coefficients are thresholded to remove noise while the approximation coefficients undergo image deblurring to enhance

facial details. Afterwards, adaptive gray level stretching is applied to improve the contrast of facial features.

6.4.1 Algorithm description

The sharpness of each probe image is computed and its value is compared with a predefined threshold S_{th}. If the current sharpness value is smaller than S_{th}, image enhancement is performed. In so doing, only images that may deteriorate the overall FRR are processed. Images with acceptable sharpness are fed to the face recognition engine directly to prevent a possible increase in image noise from unnecessary enhancement. The importance of choosing an efficient measure of face image quality becomes evident. Another advantage of using a face quality measure is attributed to the reduced computational complexity, which is also crucial to real-time applications. The block diagram of the proposed algorithm is depicted in Figure 6.13.

1. Compute the sharpness measure of the input face image, S.
2. If $S < S_{th}$, go to step 3. Otherwise, go to step 1 and wait for the next probe.
3. Decompose the face image via the Haar wavelet transform of level one.
4. Apply deblurring algorithms to the approximation image and denoising algorithms to the vertical/horizontal/diagonal detail images.
5. Apply adaptive grey level contrast stretching.
6. Reconstruct the image via the inverse Haar wavelet transform.

A global thresholding is applied for denoising all detail images. Two types of deblurring algorithms, unsharp masking (UM) and regularized deconvolution, are implemented to enhance the approximation image. The UM method uses the Laplacian filter while the regularized deconvolution utilizes the Lasso regularization.

Since the blurred image can be modeled as the original image convolved with a 2D blurring filter, the goal of image deconvolution is to undo the process and in turn eliminate the blur. Image deconvolution, as a typical inverse problem, is ill-posed.

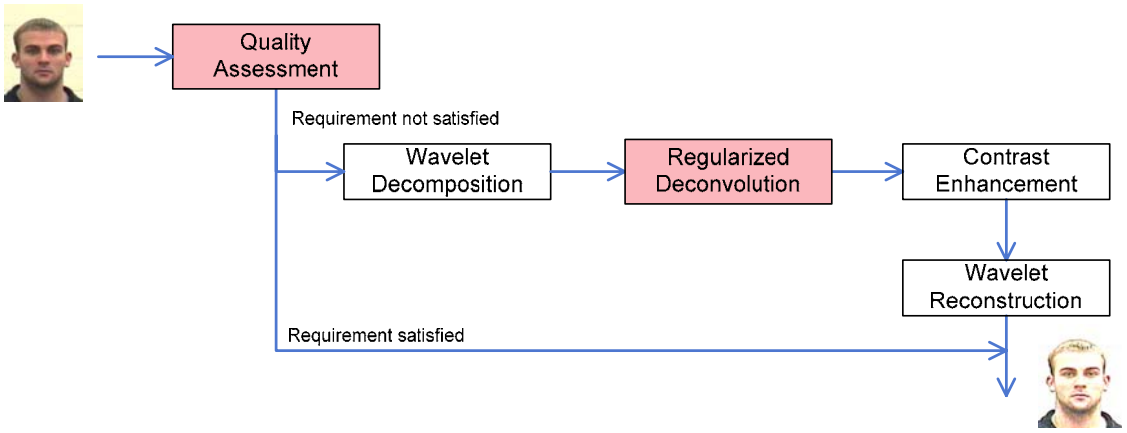


Figure 6.13. Block diagram of the enhancement algorithm for long range and high magnification face images.

Regularization is a popular approach to solve the problem where an additional term describing the smoothness of the solution is added. Without regularization, image noise sometimes would be severely amplified to where the output image is overwhelmed by noise. A typical regularized deconvolution solves the following minimization problem:

$$\mathbf{f}_\lambda = \arg\min\left\{\| B\mathbf{f} - \mathbf{f}_b \|_{L_2}^2 + \lambda_r \| L\mathbf{f} \|_{L_2}^2 \right\}, \tag{6.8}$$

where $\mathbf{f}$ and $\mathbf{f}_b$ are the original and blurred images in vector format, B and L represent the blurring filter and a predefined mask in vector format, and λ_r denotes the regularization parameter. Various forms of L can be found in literature, among which the identity matrix and Laplacian filter are two popular choices [Tikhonov77]. The Tikhonov regularization uses a norm-2 definition, which does not allow discontinuities in the solution and leads to overall smoothed edges in the restored images. The total variation regularization is proposed to preserve edges in the reconstructed images [Chan99]. It adopts a norm-1 definition [Agarwal07]:

$$\mathbf{f}_\lambda = \arg\min\left\{ \| B\mathbf{f} - \mathbf{f}_b \|_{L_2}^2 + \lambda_r \| \sqrt{\mathbf{f}_x^2 + \mathbf{f}_y^2} \|_{L_1} \right\}, \tag{6.9}$$

where $\mathbf{f}_x$ and $\mathbf{f}_y$ denote the vertical and horizontal image gradients in vector format. The total variation regularization is capable of preserving edges but suffers from significantly increased computational complexity. In our implementation, we utilize the Lasso regularization and design the regularization term as [Agarwal07]:

$$\mathbf{f}_\lambda = \arg\min\left\{\| B\mathbf{f} - \mathbf{f}_b \|_{L_2}^2 + \lambda_r \| \mathbf{f} \|_{L_1} \right\}. \tag{6.10}$$

The Lasso regularization achieves similar edge preservation as the total variation regularization with substantially reduced computational complexity.

6.4.2 Experimental results

Still images are used in the following experiments to exclude blurs from other sources such as subject motion, camera zooming, and improper focus. Different probe sets are obtained by processing the same image set via various enhancement methods, including UM, regularized deconvolution, Liao and Lin's Eigen-face [Liao05], and our wavelet based methods. In addition, two probe sets, the unprocessed face images and the 1× reference face images, are also included and their performances serve as comparison references. The same experiments are repeated for image sets at different magnifications and observation distances.

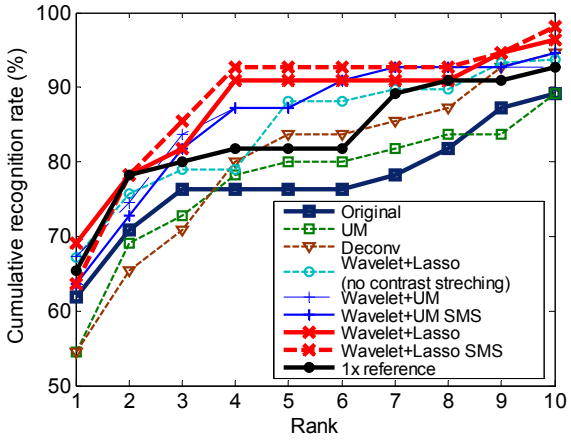
Before continuing with our discussion, we define the following notations. The probe set with face images taken at a magnification, *Mag*, a distance, *Dist* in meters, and with an inter-ocular distance of *Count* pixels, is denoted as *Mag*×*Dist*m*Count*p. Since similar observations are obtained, in the interest of space, only the comparisons based on the 10×9m60p, 20×16m60p, 109×100m80p, and 284×300m80p image sets are illustrated.

For the 20×16m60p data set, as shown in Figure 6.14 and Table 6.4, wavelet based methods are able to achieve the most improvement with an increase of 14.8% and 15.2% in CMCM for the UM and Lasso regularized deconvolution approaches yielding a performance comparable to the 1× reference. With proper processing, the degradation in FRR caused by magnification blur can be successfully compensated for. Compared with the UM based approach, the Lasso regularized deconvolution method presents a better performance. Considering the increased computations required by image deconvolution, the Lasso regularized deconvolution is well suited for applications placing more emphasis on accuracy, while the UM based algorithm achieves a better balance between accuracy and computation complexity.

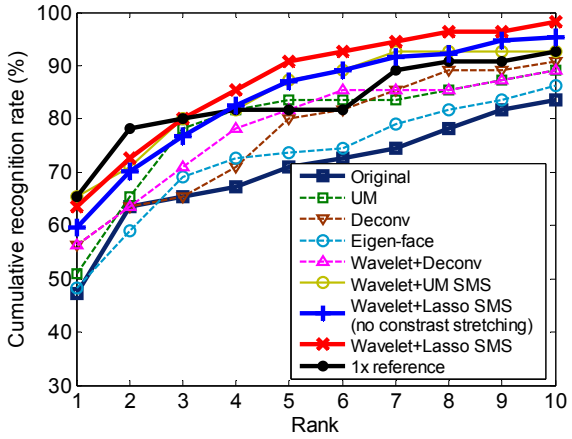
In this work, we want to use sharpness measures to predict FRR at different system magnifications and determine whether further enhancement is necessary. With sharpness measure selection (SMS) based on the threshold derived from Figure 6.11, 3.6% of the samples from the 20×16m60p image set meet the minimum criterion and hence require no further processing. The resulting performance is identical to the case where all images are processed, which verifies the suitability of the derived threshold.

For the outdoor data sets shown in Figure 6.15, our enhancement algorithm can improve the rank-one recognition rate from 46.8% to 61.1% for the 109×100m80p data set and from 29.8% to 36.9% for the 284×300m80p data set. As the system magnification increases, the improvement in FRR decreases. Different from the indoor session, where a similar FRR is achieved as the 1× reference after image quality assessment and enhancement, the performance gap between the 1× reference and the high magnification data sets remains for the outdoor session, especially for data sets with a system magnification beyond 100×. The outdoor images experience nonuniform magnification blur due to air turbulence. In our current algorithm, a uniform point spread function is estimated and used to deblur the whole image. To overcome the degradations from nonuniform blur, the point spread function should be adaptively estimated according to different regions within one image.

(a) (b) (c) (d) (e)



(f)

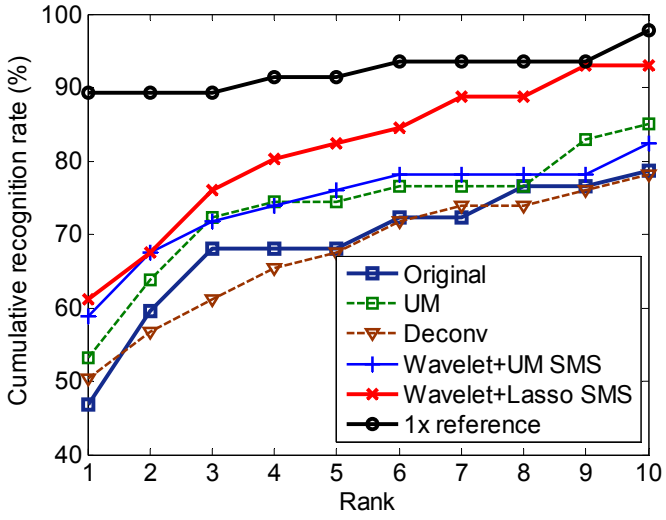


(g)

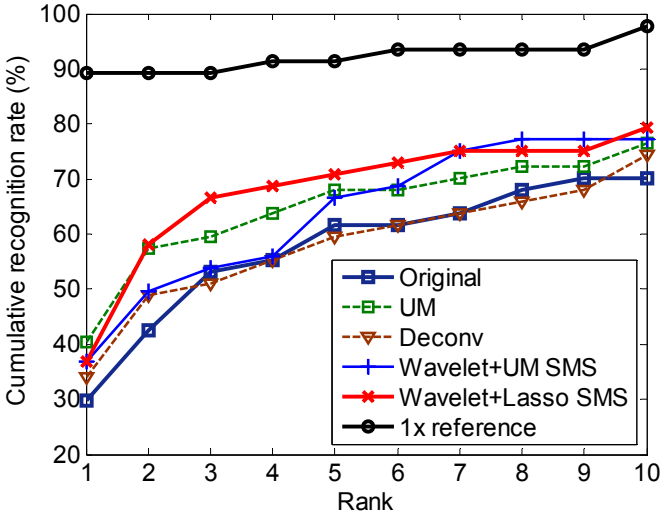
Figure 6.14. Sample images from the 20×16m60p set: (a) original image, (b) enhanced by UM, (c) enhanced by wavelet transform with the approximation image processed by UM, (d) enhanced by wavelet transform with the approximation image processed by Lasso regularized deconvolution. (e) 1× reference image. CMC comparison across probes processed by different enhancement algorithms for the indoor data sets (f) 10×9m60p and (g) 20×16m60p. The performances of the wavelet Lasso/UM algorithm with and without SMS are identical for the 20×16m60p data set. Only the CMC curves of wavelet Lasso/UM SMS are shown in (g).

Table 6.4. Performance comparison of CMCM and rank-one recognition rate across probes processed by different enhancement algorithms.

Indoor				
	10×9m60p		**20×16m60p**	
Probe set	CMCM (%)	Rank-one (%)	CMCM (%)	Rank-one (%)
Original	**69.7**	**61.8**	**58.8**	**47.3**
Eigen-face	59.7	48.2	59.7	48.2
UM	65.8	54.5	64.3	50.9
Deconv	66.1	54.5	65.3	56.4
Wavelet + Deconv	65.6	52.7	66.0	56.4
Wavelet + Lasso SMS (no contrast stretching)	75.7	67.2	70.6	59.7
Wavelet + UM	75.7	65.5	73.6	65.5
Wavelet + UM SMS	73.6	63.6	73.6	65.5
Wavelet + Lasso	75.7	63.6	**74.0**	**63.6**
Wavelet + Lasso SMS	**77.7**	**69.1**	**74.0**	**63.6**
1× reference	**74.3**	**65.5**	**74.3**	**65.5**
Outdoor				
	109×100m80p		**284×300m80p**	
Probe set	CMCM (%)	Rank-one (%)	CMCM (%)	Rank-one (%)
Original	**57.5**	**46.8**	**42.6**	**29.8**
UM	63.2	53.2	52.6	40.4
Deconv	58.0	50.4	45.9	34.0
Wavelet + UM SMS	66.4	59.0	48.9	36.9
Wavelet + Lasso SMS	**70.0**	**61.1**	**52.4**	**36.9**
1× reference	**90.7**	**89.4**	**90.7**	**89.4**



(a)



(b)

Figure 6.15. CMC comparison across probes processed by different enhancement algorithms for the outdoor data sets: (a) 109×100m80p and (b) 284×300m80p.

7 Auto-focusing for high magnification imaging

Auto-focusing is an indispensable function for imaging systems used in surveillance. For our high magnification imaging system to be useful in real-time tracking scenarios, it is critical to keep the moving target in focus. In a composite imaging system, the focus of the scope plays the dominant role. Although digital cameras are equipped with auto-focusing, scopes are available only with manual focus control. To facilitate remote and automatic control of such high magnification imaging systems, the auto-focusing capability is to be integrated.

The remainder of this chapter is organized as follows. The setup of our high magnification imaging system is described in section 7.1. A brief review of existing auto-focusing algorithms along with a performance comparison is given in section 7.2. Our auto-focusing algorithm designed for high magnification imaging systems is presented in section 7.3.

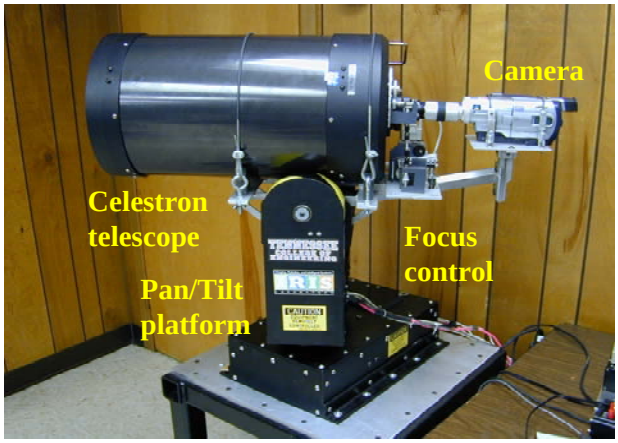
7.1 System setup

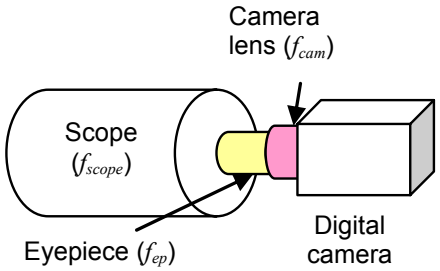
Our high magnification imaging system, equipped with high speed and remote pan/tilt/focus control, is shown in Figure 7.1. To fully explore the optical capabilities of the Celestron scope and the Sony camcorder, an afocal coupling is selected. The Celestron scope is connected to the Sony camcorder via a Celestron Plössl eyepiece. The focal lengths of the Celestron scope and eyepiece are 2800mm and 40mm, respectively. The Sony camcorder has a 47mm~846mm zoom capability.

The scope magnification is defined as:

$$M_{scope} = f_{scope} / f_{ep}, \tag{7.1}$$

where f_{scope} and f_{ep} denote the focal lengths of the scope and the eyepiece, respectively. For an afocal coupling, the system magnification M_{sys} is the product of the scope magnification M_{scope} and the camera's normalized magnification M_{cam} given by:





(a) (b)

Figure 7.1. (a) System setup and (b) illustration of the afocal coupling for composite imaging systems. Fully motorized pan/tilt/zoom and auto-focusing capabilities facilitate remote and automatic control. The resulting system can perform object tracking and monitoring in the same fashion as commercial PTZ cameras.

$$M_{cam} = f_{cam}(mm)/50\,, \tag{7.2}$$

where f_{cam} is the camera's focal length expressed in the 35mm equivalent standard. Based on the focal length specification of each component, the achievable system magnification is approximately 70× to 1200×.

The Celestron scope's existing focus control features a manually operated knob requiring 40 full turns to cover the complete focus range. To automate it, we coupled the control to an Animatics SmartMotor through a gear drive of our own design. The main requirement is that the system be precise enough to give repeatable control positioning with increments as fine as the smallest resolution that starts to produce noticeable degradation in the resulting images. The empirical minimum resolution is found to be less than 40 degrees of knob rotation. When converted to motor steps and normalized to the minimum resolution, the dynamic focus range is -200 to 200 steps.

7.2 Algorithm comparison

7.2.1 Algorithm review

In literature, there exist two main groups of auto-focusing methods: active and passive. In active auto-focusing, range finding sensors are used to determine the distance between the camera and the target. Passive auto-focusing can be further divided into two categories: device based and image based. Device-based passive auto-focusing employs additional devices, such as a split prism. The image-based approach requires no extra

equipment. The optimal focus is found by evaluating a sequence of images collected at various focus positions.

Estimating depth from defocused images is one major direction in image-based passive auto-focusing algorithms [Subbarao92]. The distance between the camera and the target is estimated from several defocused images. The degree of blur in the images is characterized as the variance of a Gaussian kernel. Target depth is expressed as a function of these variances and can be computed when these variances are available. In our application, a considerable amount of blur comes from high magnification rather than improper focus. The simple relation between blur and depth is not entirely valid. For this reason, the use of this type of methods in high magnification systems remains questionable.

In the second main branch of image-based passive auto-focusing algorithms, the optimal focus is found by searching for the focus location that yields an image with the highest sharpness value. Various search strategies have been developed. The Fibonacci search is the best-known algorithm [Krotkov89], which guarantees that the maximum of the criterion function is found within a known number of iterations depending only on the focus range. The hill-climbing search divides the procedure into two stages: out-of-focus region (coarse) search and focused region (fine) search. Given a heuristic choice of step magnitudes, the hill-climbing search is able to converge to the optimal focus. A number of hill-climbing algorithms have been proposed with modifications regarding the selection of step sizes, termination criteria, the size of the search window, *etc* [Choi99, He03, Ooi90].

Variations are introduced to these basic algorithms for a better performance. For instance, in the fine search stage, the image sharpness is evaluated at three focus locations and these samples are fitted to a quadratic or a Gaussian function, the maximum of which is the estimated focused position [Subbarao98]. Lee *et al*. employed different sharpness measures for the coarse and fine search stages [Lee95]. In the coarse search stage, measures with low computational cost and low sensitivity to sidelobes, such as variance based measures, are used. Gradient based measures, for instance the Tenengrad measure, are used for the fine search. To avoid the back-and-forth motor motion required by the Fibonacci search, Kehtarnavaz and Oh proposed a sequential search algorithm, referred to as the rule-based search, where the step size is varied according to the distance from the best focus location [Kehtarnavaz03].

Special patterns, such as a radial test pattern, are also employed to calibrate the best focus position for applications with a fixed distance between the target and the camera [Lin03]. Since the image with the best focus should have the smallest blurred region and hence the smallest equivalent radius, Lin *et al*. applied the circular Hough transform to determine the radius of the center blurred image, and from this obtained the best focus position.

7.2.2 Experimental results

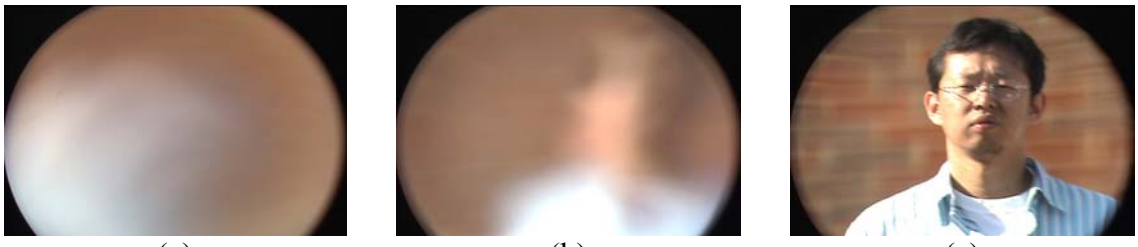
To evaluate the performance of various search algorithms, each in conjunction with different sharpness measures, we carried out the following experiments. Images are

collected at uniformly distributed focus positions and their sharpness measures are computed. A search algorithm is then applied to locate the best focus position. Ideally, the estimated focused position should correspond to the maximum sharpness value. Any difference (expressed in motor steps) between them is the estimation error, the size of which is translated into the accuracy of the search algorithm. Another performance criterion, the speed of convergence, is described by the number of iterations and the number of motor steps traveled before the optimal focus is obtained. These two factors (iterations and motor steps) are often closely related, with a large number of iterations resulting in a large number of motor steps. An exception is the Fibonacci search, where a small number of iterations is guaranteed, but where a large number of motor steps often results from the algorithm's back-and-forth search behavior.

Four low magnification image sequences (2.28×), resolution chart (RC), Hello-Kitty doll (HD), license plate (LP), and man's face (MFL), are collected by the Canon A80 camera at an interval of three focus motor steps covering a focus range from 0.2m to infinity with a total of approximately 60 images per sequence. The RC and LP sequences exemplify images with strong and clustered edges. Ten high magnification sequences are collected by the Sony TVR730 camcorder and the Celestron scope. Various system magnifications are used: 70×, 100×, 245×, 500×, and 1500×. At each magnification, two sequences (400 frames per sequence) are collected, one of a scene with strong and clustered edges of a brick wall (BW) and the other with scattered and low contrast edges of a man's face (MFH). Figures 7.2 and 7.3 show sample images from the LP and MFH (70×) sequences collected at the best focus position and at the end points of the focus

(a) (b) (c)

Figure 7.2. Sample images from the LP sequence (system magnification: 2.28×, target distance: 1m): (a) far focus end, (b) near focus end, and (c) best focus.



(a) (b) (c)

Figure 7.3. Sample images from the MFH sequence (system magnification: 70×, target distance: 65m): (a) far focus end, (b) near focus end, and (c) best focus.

range.

We limit our experiments to four types of sharpness measures excluding the statistics based measures due to their inferior performance. Table 7.1 lists the sharpness measures used. Various search algorithms are implemented, including the binary search (BS), Fibonacci search (FS), and rule-based search (RS). In addition, quadratic function fitting is applied to the fine search stage, following the coarse search based on the binary search and the Fibonacci search. The resulting algorithms are referred to as BF and FF, respectively. Also implemented is one example of the hill-climbing search (HC) [Choi99].

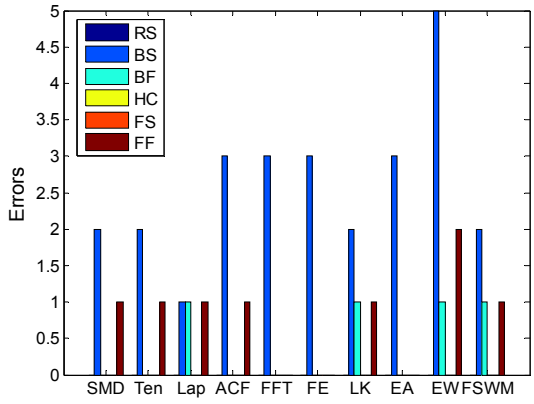
Figure 7.4 studies various search algorithms based on the errors in the detected focus position, the number of iterations, and the number of motor steps using the LP sequence. In terms of accuracy, the Fibonacci, hill-climbing, and rule-based searches produce the best performance. However, the performance of the hill-climbing search is sensitive to the parameters used. These parameters must be selected carefully, especially for noisy applications.

With the Fibonacci search, the number of iterations is fixed for a given focus range. However, the Fibonacci search involves the most back-and-forth motions and therefore the most motor steps. Although it needs a similar number of iterations as the binary search, the rule-based search involves only unidirectional movements and hence requires fewer motor steps. The use of function approximation avoids unnecessary iterations during the fine search stage, thereby reducing the total number of iterations and motor steps.

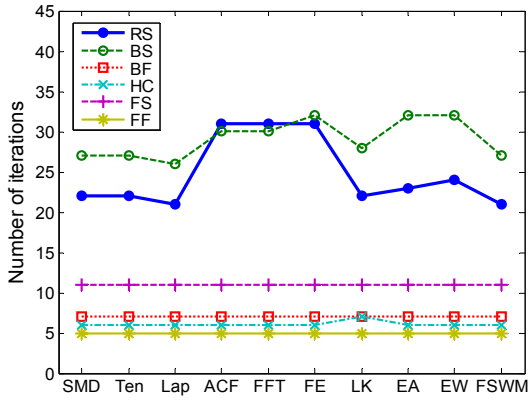
Figure 7.5 demonstrates the experimental results using high magnification sequences. Due to magnification blur, the computed sharpness measures are noisy, leading to obviously increased estimation errors. The binary search and the hill-climbing search, inherently sensitive to image noise and magnification blur, present the most performance degradation.

Table 7.1. Sharpness measures used in the comparison of auto-focusing algorithms.

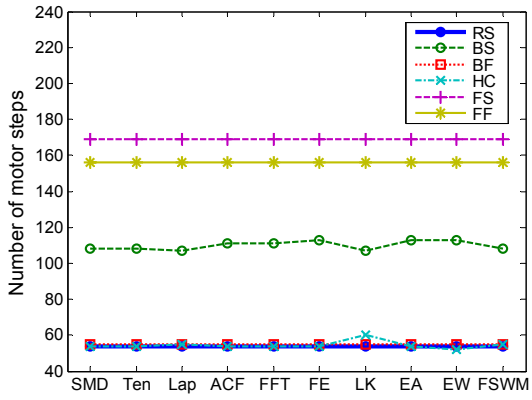
Category	Sharpness measure	Reference
Gradient based	Sum-Modulus-Difference (SMD)	[Santos97]
	Tenengrad (Ten)	[Kroktov89]
	Laplacian (Lap)	[Kroktov89]
	Frequency selective weighted median (FSWM)	[Choi99]
Autocorrelation based	Area of the central peak of the autocorrelation function (ACF)	[Batten00]
Transform based	Fast Fourier transform (FFT)	[Subbarao92]
	Frequency entropy (FE)	[Kristan04]
Edge based	Edge width (EW)	[Li02]
	Edge area (EA)	[Dijk02]
	Local kurtosis (LK)	[Caviedes02]



(a)

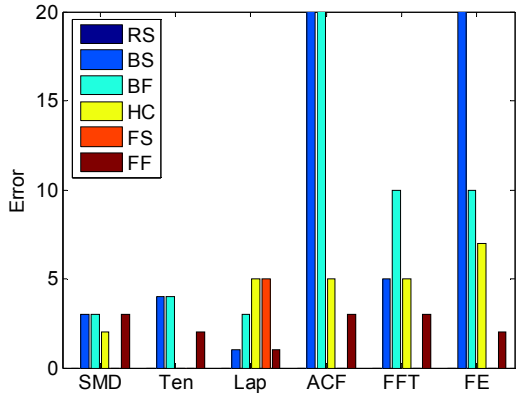


(b)

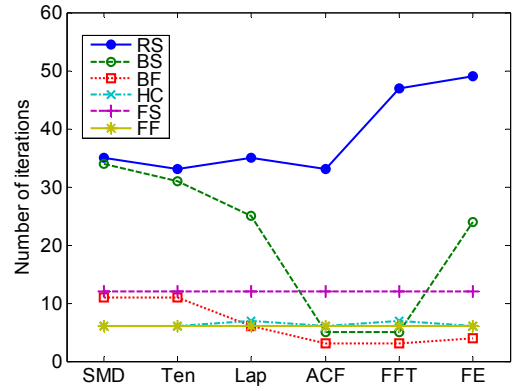


(c)

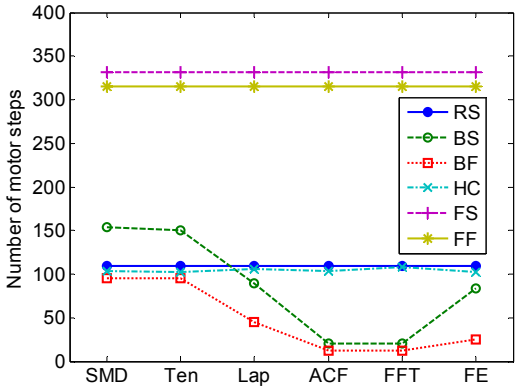
Figure 7.4. Performance comparison of various search algorithms in conjunction with various sharpness measures using the LP sequence. (a) Estimation error expressed in motor steps (the estimation errors for RS, HC, and FS are zero). (b) The total number of iterations used before obtaining the optimal focus position (the smallest number of iterations: FF and HC). (c) The total number of motor steps traveled before obtaining the optimal focus position (the smallest number of motor steps: RS, BF and HC).



(a)



(b)



(c)

Figure 7.5. Performance comparison of various search algorithms in conjunction with various sharpness measures using the MFH sequence (system magnification: 70×, target distance: 65m). (a) Estimation error expressed in motor steps (the largest performance degradation: BS, BF, and HC). (b) The total number of iterations used before obtaining the optimal focus position (the smallest number of iterations: FF and HC). (c) The total number of motor steps traveled before obtaining the optimal focus position (the smallest number of motor steps: RS, BF and HC).

Similar behaviors are observed for the remaining image sequences. As a conclusion, we compare the tested search algorithms based on three criteria: accuracy, speed of convergence described by the number of iterations and motor steps, and stability (sensitivity to image noise, parameter selection, and magnification blur). Overall, the rule-based search and the Fibonacci search with function fitting generate the best performance. In our real-time auto-focusing system, extra attention is paid to the number of motor steps, since our system has a larger focus range compared with low magnification imaging systems. Therefore, the rule-based search is the most promising method, which falls in the category of sequential search with variable step sizes.

7.3 Auto-focusing for high magnification imaging systems

To design the auto-focusing algorithm for high magnification imaging systems, we experience two major difficulties. (1) For a large visible distance, our high magnification imaging system involves a large focus range varying from 20m up to 1000m (infinity). (2) The collected images suffer substantially from degradations such as increased image noise and severe image blur caused by high magnification and air turbulence, producing time varying and noisy sharpness measures. These two difficulties impose additional requirements on the design of a proper auto-focusing algorithm, especially the speed of convergence and robustness to image degradations.

In light of the system limitations – *i.e*. wide focus range and noisy sharpness measures – and the performance comparison of various search algorithms [Yao06F], sequential search algorithms with variable step sizes are selected. The sequential search completes peak detection in one sweep, nearly eliminates changes in motion direction, and saves on motor steps. Variable step size optimizes the motor step distribution and minimizes the number of iterations. The remaining questions are: (1) when and how to change the step size and (2) how to evaluate image sharpness appropriately. The derivation of transition criteria and the selection of sharpness measures answer the above questions, respectively.

7.3.1 Transition criteria

The step sizes are adjusted adaptively throughout the search process according to the current focus location. The small, medium, and large step sizes are used in the peak, ramp, and saturation regions, respectively. From the viewpoint of a state transition machine, three distinctive states can be defined. The state transition representation associates the search process with an estimation process, where the optimal sequence of state transitions is retrieved given a sequence of noisy observations and a predefined structure (states and transition hypothesis). Consequently, maximum *a posteriori* and maximum likelihood estimation can be applied. Most of the sequential search algorithms use empirical thresholds to govern the step size transitions. Based on the state transition

representation, these thresholds can be indeed derived from maximum likelihood estimation.

To build probabilistic models for state transitions, the statistical behavior of sharpness measures is studied. The search process is divided into two stages: the pre-peak stage where no peak is detected and the post-peak stage where a possible peak is detected. In the pre-peak stage, the determinant variable is ΔS, the difference between consecutive sharpness measures, while in the post-peak stage, the focus is shifted to the absolute value of the image sharpness S. In our implementation, $S_{\max}$, the recorded maximum sharpness value, is used as a reference. We examine the statistical behavior of $\Delta S / S_{\max}$ and $S/S_{\max}$ and obtain the thresholds assuming that both variables obey a Gaussian distribution. In practice, to avoid back-and-forth switches caused by noise, some state transitions are issued only when the corresponding transition criteria are satisfied three times. The following counters, C_{down} and C_{flat}, are defined for the ramp region in the post-peak stage and the saturation region, respectively. Table 7.2 summarizes the transition criteria. Assuming that the current state is peak and $\Delta S < 0$, C_{down} increases by one. The consecutive state is ramp if C_{down} is larger than or equal to three and remains in peak otherwise.

7.3.2 Sharpness measure selection

The proper use of sharpness measures is also of great importance to the system performance. Sharpness measures respond to the changes in camera focus in quite different ways. Variance based sharpness measures produce gradual slopes while gradient based sharpness measures produce sharp peaks [Lee95]. However, the performance of variance based sharpness measures deteriorates for high magnification images. In some cases, they could not even preserve the desired unimodal shape as an appropriate sharpness measure.

From the analysis of their properties, we observe that autocorrelation based measures (ACF) generate responses with varying slopes depending on the window size used, as shown in Figure 7.6. Measures with a larger window size produce wide peaks and

Table 7.2. Transition criteria. Assuming that the current state is *peak* and $\Delta S < 0$, C_{down} increases by one. If C_{down} is larger than or equal to three, the next state is *ramp*. Otherwise, remain in *peak*.

<table>
<tr><th colspan="2" rowspan="2"></th><th colspan="3">End state</th></tr>
<tr><th>Peak</th><th>Ramp</th><th>Saturation</th></tr>
<tr><td rowspan="4">Start state</td><td rowspan="2">Peak</td><td colspan="2">$\Delta S < 0,\ C_{down}++$</td><td rowspan="2">$S \leq 0.24 S_{max}$</td></tr>
<tr><td>$C_{down} < 3$</td><td>$C_{down} \geq 3$</td></tr>
<tr><td rowspan="2">Ramp</td><td rowspan="2">$\Delta S > 0.23 S_{max}$</td><td colspan="2">$\Delta S > 0.09 S_{max}$, $C_{flat}++$</td></tr>
<tr><td>$C_{flat} < 3$</td><td>$S \leq 0.24 S_{max}$ or $C_{flat} \geq 3$</td></tr>
<tr><td>Saturation</td><td>$\Delta S > 0.23 S_{max}$</td><td>$\Delta S > 0.09 S_{max}$</td><td>Otherwise</td></tr>
</table>

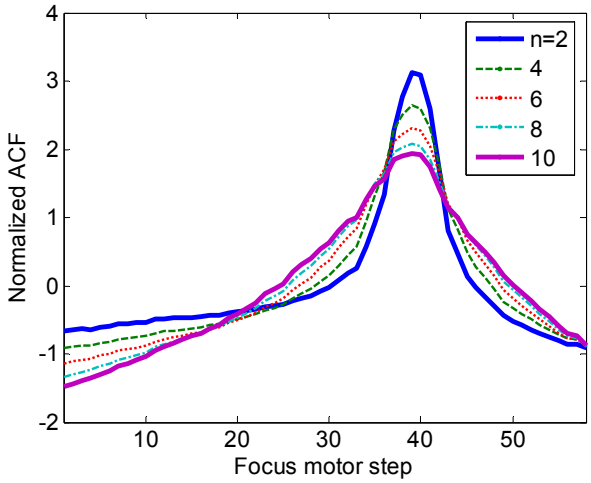


Figure 7.6. ACF sharpness measure with various window sizes for the LP sequence. *n* denotes the widow size.

gradual slopes. Therefore, the ACF measure with a large window size can be used in the coarse search stage, and gradient based sharpness measures can be used in the fine search stage.

In practice, the combination of two types of sharpness measures is used to improve the response shape and suppress noise. The summation of the Tenengrad and ACF with $n = 10$ (ACF_{10}) produces an improved slope in the ramp region, corrects local extrema in the responses of single measures, and reduces noise, as shown in Figure 7.7. In our application, the summation of two sharpness measures is used for imaging systems with higher magnifications (250×~500×).

7.3.3 Experimental results

Experiments based on offline image sequences (indoor/outdoor, low/high magnifications) are conducted. Three types of sharpness measures are used: gradient based, autocorrelation based, and frequency domain based. In terms of accuracy, speed of convergence, and resistance to image noise and blur, the rule-based search and the Fibonacci search with function fitting outperform other search algorithms. Therefore, these two methods are selected as comparison references. In the interest of space, only the experimental results for the MFH sequence with a magnification of 70× are presented in Figure 7.8.

From Figure 7.8, our algorithm achieves an accuracy comparable to the RS algorithm. Meanwhile, our algorithm requires a smaller number of iterations compared with the RS algorithm and the lowest number of motor steps. Overall, our algorithm provides a better balance between accuracy and the speed of convergence.

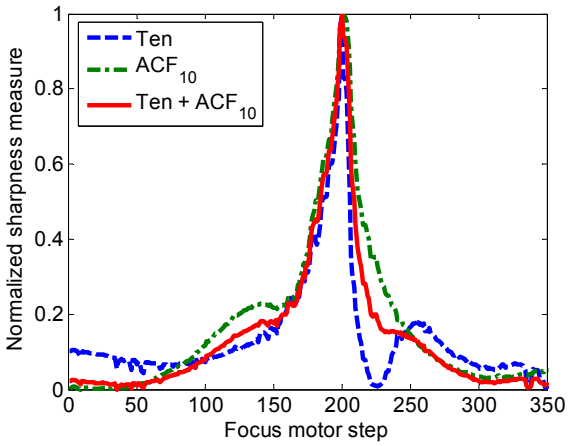


Figure 7.7. Comparison of the Tenengrad (Ten) measure, ACF measure, and a linear combination of these two measures for the MFH (100×) sequence.

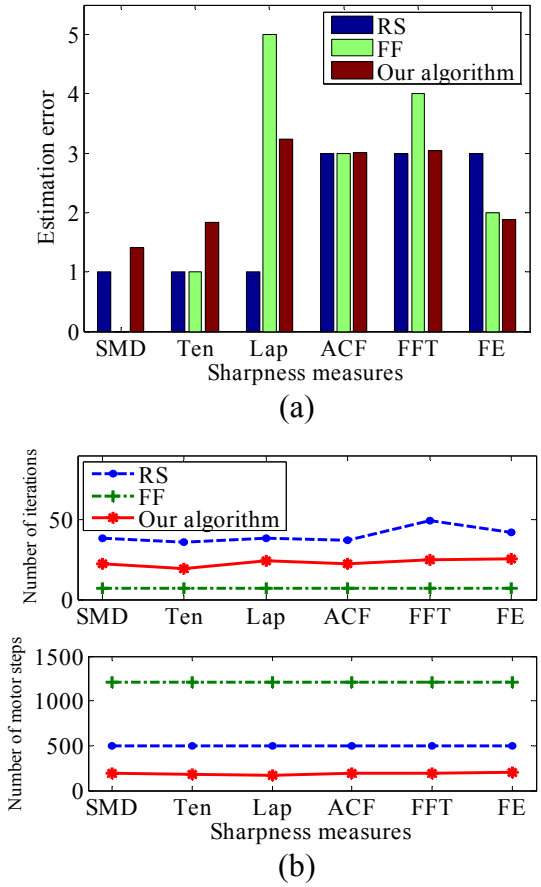
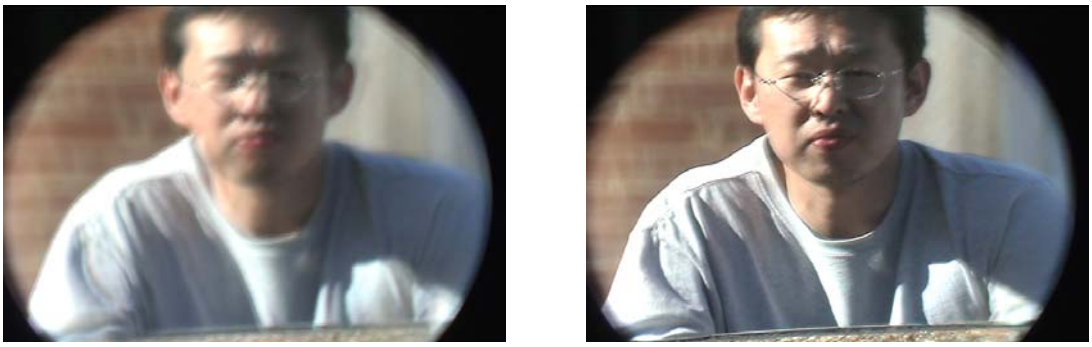


Figure 7.8. Comparison across sharpness measures and search algorithms including RS, FF, and our auto-focusing algorithm at 70× magnification. (a) Estimation error. (b) The total number of iterations and motor steps.

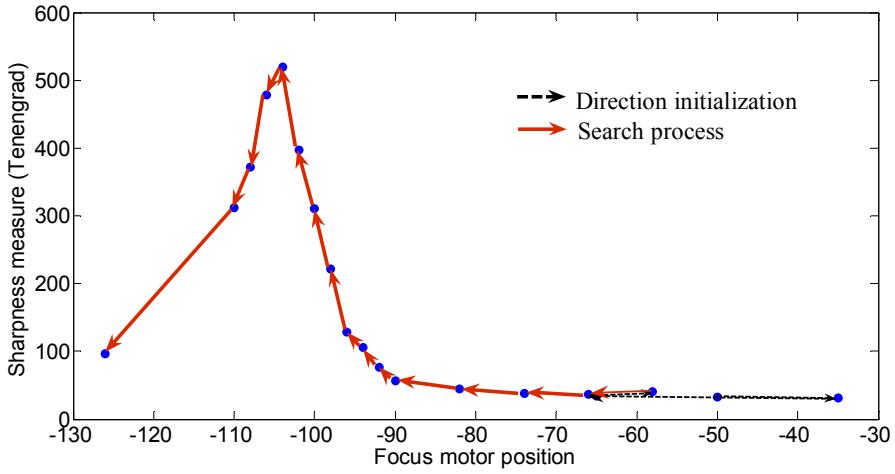
Figures 7.9 and 7.10 show the sampled images in two real-time auto-focusing sequences collected at a system magnification of 70× and 500×, respectively. Figures 7.9(e) and 7.10(e) illustrate the sampled focus positions and their sharpness measures. Given a starting point within ±100 motor steps from the peak region, and with a frame rate of approximately 7.2 frames per second, our algorithm can precisely detect the optimal focus position in 2 seconds.

Based on the raw input images, our auto-focusing algorithm works well for a system magnification up to 250×. Further increases in magnification result in severely blurred images which undermine the ability of the sharpness measures to produce smooth and unimodal curves. Image pre-processing and the use of a summation of two types of sharpness measures are possible solutions.

(a) (b)

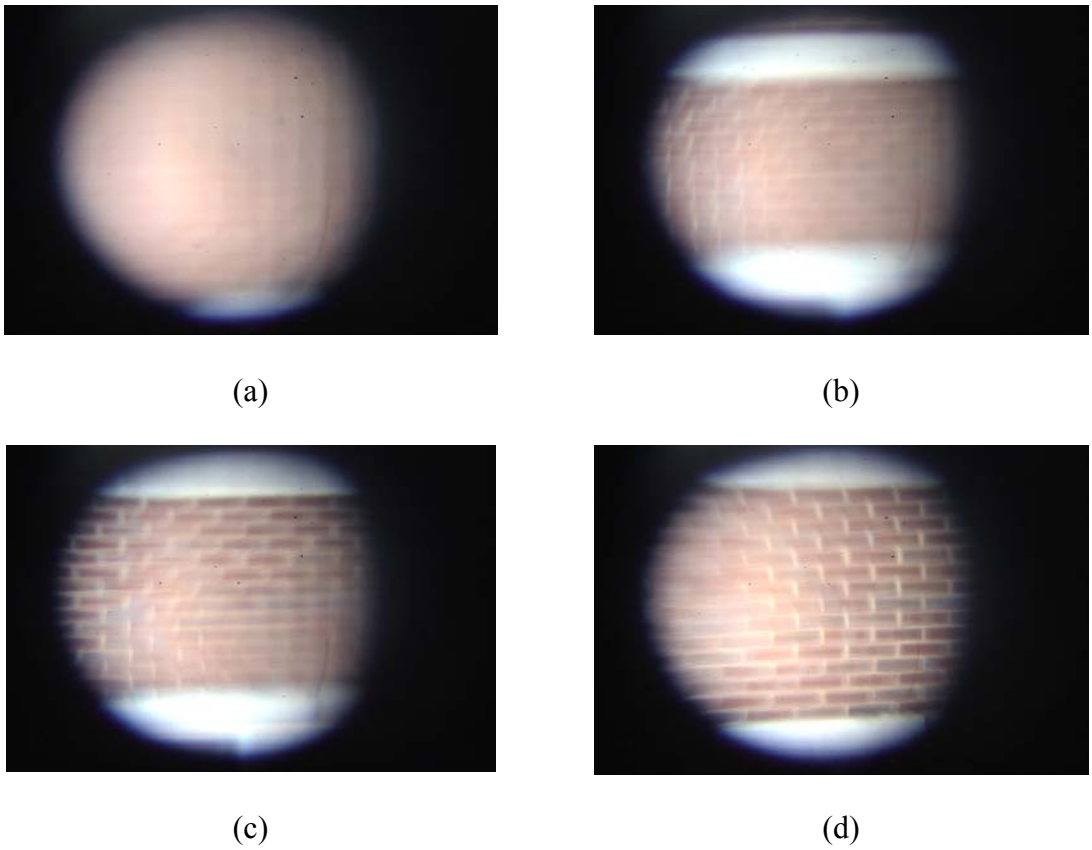


(c) (d)



(e)

Figure 7.9. Auto-focusing for the MFH sequence (magnification: 70×, distance: 65m). Sample frames collected at: (a) initial focus position, (b) intermediate focus position, (c) last evaluated focus position, and (d) best focus position. (e) Sampled focus positions. Starting position: -50. Estimated optimal focus position: -102. Motor steps: 106. Time: 1.9s.



(e)

Figure 7.10. Auto-focusing for the BW sequence (magnification: 500×, distance: 300m). Sample images collected at: (a) initial focus position, (b) intermediate focus position, (c) last evaluated focus position, and (d) best focus position. (e) Sampled focus positions. Starting point: 0. Estimated optimal focus position: -28. Motor steps: 96. Time: 1.8s.

8 Conclusions

In this dissertation, we have addressed two key issues in successfully establishing an automatic surveillance system: coverage and resolution. Sensor planning algorithms were proposed to resolve the coverage issue. High magnification imaging was introduced to achieve the required resolution and size preserving tracking was utilized to maintain the required resolution. The collaboration between high magnification imaging and size preserving tracking makes long range surveillance from hundreds of meters feasible. In previous chapters, we have presented a survey of multi-camera surveillance systems, derived our theoretical framework, and demonstrated the effectiveness of the proposed methods via extensive experiments and comparisons with existing leading algorithms. We conclude this dissertation with a brief summary of the contributions and a short discussion of the directions for future research.

8.1 Summary of contributions

Algorithms regarding sensor planning and size preserving tracking are proposed for automated and persistent surveillance. For long range applications, high magnification imaging systems with automated image quality assessment and enhancement are employed. The key contributions of this research are the following.

- **Sensor planning**: the proposed sensor planning method improves existing algorithms by adding handoff rate analysis for environments with multiple dynamic targets. The optimal balance between the overall coverage and handoff success rate is achieved.

- **Size preserving tracking:** the size preserving tracking algorithm with linear computations is developed based on the more advanced paraperspective projection model producing improved estimation accuracy and robustness to disturbances from practical tracking, such as image noise and system latency.

- **Quality assessment and enhancement of high magnification images**: adaptive sharpness measures capable of suppressing artificial responses from image noise are designed for the evaluation of image quality under high magnifications.

Wavelet based enhancement algorithms with automated frame selection capability are developed to strengthen facial features for an improved face recognition rate.

For each of these contributions, we have presented both quantitative and quality comparisons to demonstrate their strength and analyze their limitations. We have submitted the work regarding sensor planning to [Yao08A] for review. A dual camera system as an implementation of multi-camera surveillance was established and generated publications in [Yao06E, Yao07E]. We have presented the size preserving tracking algorithms in [Yao06B, Yao06C]. A journal version of the work is under review [Yao07D]. Several publications have resulted from our work regarding high magnification imaging, including papers focusing on system design [Yao06A, Yao07A], a paper describing our high magnification face database [Yao06D], and a book chapter and a journal paper discussing our enhancement algorithm [Yao07B, Yao07C].

8.2 Cost analysis

Considering the volume of the development conducted in this dissertation, two fair questions arise: (1) how much did it cost to develop this technology and (2) how much would it cost in 2008 US dollars to implement a typical system. A typical system, for instance, would consist of a room with dimensions of 15m×20m needing two or three PTZ cameras and two or three omnidirectional cameras to monitor the entire environment.

The initial development for the generic sensor placement and size preserving tracking algorithms took the equivalent of three years of a full time Ph.D. work. Student stipend, tuition, supervision cost of the lead and associated faculty, and overall equipment requirements average approximately $80,000 per year. A more detailed cost analysis for one calendar year is listed in Table 8.1. Therefore, in terms of sheer time, this activity necessitated over $250,000 worth of funding. Added to this is the prior experience of the student and faculty in generating new ideas for solving these difficult problems. Therefore, the theoretical development and software implementation, testing, and validation constitute an initial total cost that is often times ill-estimated by many in the community.

For an industrial company or business that desires to build a fully functioning prototype for an area that is comparable to the one cited above, the estimated cost is roughly $100,000 to $200,000 depending on the complexity of the environment to monitor and the specific requirements for access control, object tracking, threat awareness, and decision making. This technology, if implemented at a large scale where 100 to 1000 copies are put into service, becomes a viable solution where the initial development cost is distributed over the total number of systems sold. It is estimated that the hardware costs approximately $30,000 and that the equivalent licensing of the software requires $20,000, hence making the total cost about $50,000 per unit. This technology is highly evolving and requires frequent maintenance and upgrade to avoid obsoleteness, which in consequence adds an additional $10,000 per year for updating. For the next few years, if not decades, access control, object tracking, and threat

Table 8.1. Itemized budget for one calendar year.

A	Personnel						
	Name	Type appt. (months)	Effort on project	Base salary	Salary requested	Fringe benefits	Total
	Professors (2)	12	11%	80,000	17,600	4,928	22,528
	Graduate student	9	100%	22,000	16,500	1,206	17,706
B	**Total Personnel**				34,100	6,134	40,234
C	Travel (domestic)						2,000
D	Maintenance and repairs						1,000
E	Supplies						1,000
F	Equipment						4,500
G	Graduate student tuition (school year rate: 2,787/semester)						11,148
H	**Total direct costs (B through G)**						59,882
I	Indirect costs / F&A		Rate	47%	Base	44,234	
	Total indirect costs / F & A						20,790
J	**Total budget**						80,672

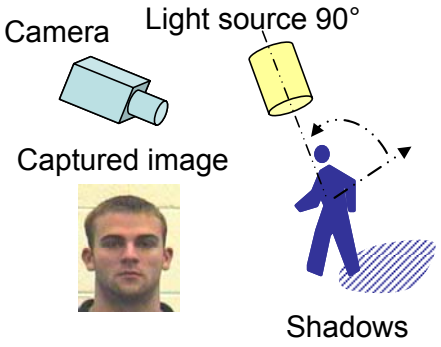
awareness will remain an active research area and a practical need that many will invest in to protect valuable assets for private industry as well as government applications.

8.3 Directions for future research

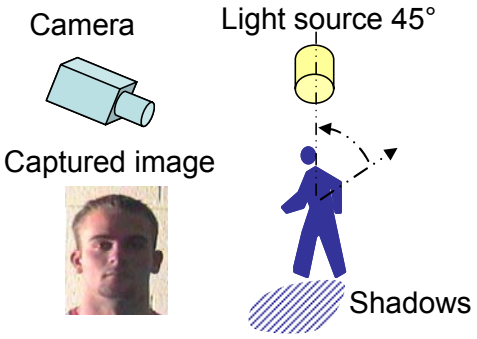
The ideas and concepts in this dissertation offer interesting avenues for future research. Although many directions are possible, we have identified the following areas as particularly important.

8.3.1 Sensor planning considering illumination

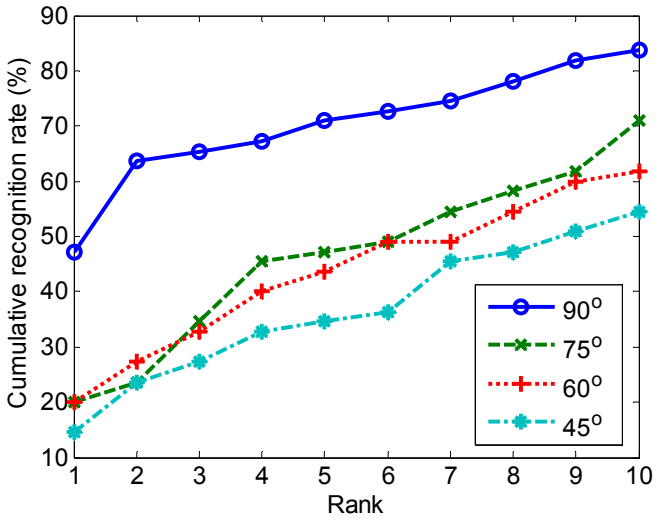
In a surveillance system, different illumination conditions cause shadows and changes in the object's appearance, which imposes considerable challenges on object tracking and recognition. Figure 8.1 shows two setups of a surveillance system with different positioning of the illumination sources. With the proper lighting in Figure 8.1(a), all facial details are visible and can be used for face recognition. In comparison, half of the facial features are poorly illuminated due to the effect of the side light in Figure 8.1(b). To illustrate the degradation in face recognition rate caused by improper illumination, image sets are collected with different lighting angles. Figure 8.1(c) compares the face recognition rate with various illumination conditions. The face recognition rate degrades significantly as the light source rotates away from the optimal position. The rank-one



(a)



(b)



(c)

Figure 8.1. The influence of the positioning of the illumination sources on the performance of face recognition. Two system setups with different positioning of the light sources: (a) 90° and (b) 45°. (c) Comparison of the face recognition rate of image sets collected under various illumination conditions.

recognition rate drops from 48.2% to 15.8% as the light source rotates from 90° to 45°. Although normalization algorithms, such as Self Quotient Image [Wang04] and multi-scale Retinex [Jobson97], can be applied to compensate for nonuniform illumination, the achievable improvement is limited [Yao07B]. Therefore, for accurate and robust surveillance, the positioning of the illumination sources is of the same significance as the positioning of the cameras.

The camera placement algorithm discussed in chapter 3 achieves the optimal balance between the overall coverage and handoff success rate excluding the effect of illumination by assuming a uniform lighting condition. One prominent future research is to relax the above assumption and take illumination into consideration. A forth term describing the illumination quality is to be added to the observation measure defined in (3.11), which depends on both the position and the characteristics of the light source. In so doing, the optimal camera placement can be derived including the effect of illumination. A further extension is to incorporate the optimization of the positions and types of the illumination sources into the search for the optimal camera placement. The final output consists of not only the optimal camera placement but also the associated positioning of the illumination sources.

8.3.2 Sensor planning considering objects with different priority ranks

It is a common practice to assign different priority ranks to objects that need to be tracked simultaneously. Given limited computational capacity, more resources are allocated to objects with higher priorities at the cost of dropping out objects with lower priorities. In section 3.2, the proposed camera placement algorithm considers the problem of camera overload based on a probabilistic framework that models multi-object tracking as a Markov chain and derives the overload probability with all the objects having the same priority. To incorporate objects with different priorities, we carry out the following derivations.

Let $N_{th,j,pr}$ denote the maximum number of objects with a priority rank less than or equal to pr that can be tracked simultaneously by the j^{th} camera. pr is in the range from 1 to N_{PR} with N_{PR} as the maximum number of priority ranks. We purposefully add $N_{th,j,0}$=0 to simplify the formulation. As the priority rank pr increases, $N_{th,j,pr}$ increases as well to ensure that more resources are allocated to objects with higher priorities. Note that $N_{th,j,N_{PR}} = N_{obj,j}$, where $N_{obj,j}$ is the maximum number of objects that can be tracked simultaneously by the j^{th} camera.

Assume that the arrival of an object with a rank pr in the FOV of the j^{th} camera follows a Poisson distribution with a rate of $\lambda_{c,j,pr}$. Its camera-residence time follows an exponential distribution at a rate of $1/\mu_{c,j,pr}$. The probability $P_{n,j}$ of the n^{th} state of the Markov chain is given by:

$$P_{n,j} = \frac{p_{o,j}}{n!} \prod_{ii=1}^{pr-1} \left(\sum_{jj=ii}^{N_{PR}} \frac{\lambda_{c,j,jj}}{\mu_{c,j,jj}}\right)^{N_{th,j,ii}-N_{th,j,ii-1}} \left(\sum_{ii=pr}^{N_{PR}} \frac{\lambda_{c,j,ii}}{\mu_{c,j,ii}}\right)^{n-N_{th,j,pr-1}}, \quad (8.1)$$

for $N_{th,j,pr-1} < n \le N_{th,j,pr}$ with

$$p_{o,j} = \left\{ \sum_{nn=0}^{N_{obj,j}} \left[\frac{1}{nn!} \prod_{ii=1}^{pr-1} \left(\sum_{jj=ii}^{N_{PR}} \frac{\lambda_{c,j,jj}}{\mu_{c,j,jj}} \right)^{N_{th,j,ii} - N_{th,j,ii-1}} \left(\sum_{ii=pr}^{N_{PR}} \frac{\lambda_{c,j,ii}}{\mu_{c,j,ii}} \right)^{nn - N_{th,j,pr-1}} \right] \right\}^{-1}. \tag{8.2}$$

The probability that the j^{th} camera reaches the maximum number of objects with a priority rank of *pr* is expressed as:

$$P_{\max,j,pr} = \sum_{n=N_{th,j,pr}}^{N_{obj,j}} P_{n,j} \,. \tag{8.3}$$

Denote the average arrival rate of an object with a rank *pr* at the i^{th} grid as $\lambda_{g,i,pr}$ and the mean camera-residence time as $1/\mu_{g,i,pr}$. The probability of camera overload at the i^{th} grid $P_{co,i,pr}$ is given by:

$$P_{co,i,pr} = (1 - e^{\lambda_{g,i,pr}/\mu_{g,i,pr}}) \prod_{j, a_{1,ij} x_j = 1} P_{\max,j,pr} \,, \tag{8.4}$$

where $a_{1,ij}$=1 if $Q_{ij} \ge Q_F$, $a_{1,ij}$=0 otherwise, and x_j=1 if the j^{th} camera is chosen. Finally the objective function used for the search of the optimal camera placement can be defined as:

$$c_i = w_1(c'_{1,i} > 0) + w_2(c'_{2,i} = 2) - w_3(c'_{3,i} > 1) + w_5 \frac{\sum_{pr=1}^{N_{PR}} pr(P_{co,i,pr} \le P_{co,th,pr})}{\sum_{pr=1}^{N_{PR}} pr}, \tag{8.5}$$

where $P_{co,th,pr}$ is a predefined threshold for priority rank *pr*. In comparison with (3.30), objects with different priority ranks are allowed and incorporated into sensor planning.

8.3.3 Sensor planning for 3D floor plans

The sensor planning algorithms presented in section 3.2 are based on 2D floor plans. A 2D floor plan is representative of environments with an approximately flat ground, where the variations along the normal direction of the ground plane (*Z* axis) are marginal. To generalize the applicability of our algorithms to arbitrary environments, the environments' 3D geometry that allows variations along the *Z* axis needs to be considered. Accordingly, the 2D mesh grid presentation of the floor plane is upgraded to a 3D mesh grid of the floor surface. Visibility analysis is carried out not only for obstacles and dynamic occlusions from moving targets but also for possible self-occlusions from the varying elevation of the floor surface.

8.3.4 Constrained deblurring of high magnification images

To improve the performance of the lasso regularized deconvolution, additional constraints can be incorporated. The non-negativity constraint is a popular choice. In deconvolution, although the intensities of the observed image are all positive, the deblurred image may contain negative values if the non-negativity constraint is not imposed. Therefore, reinforcing non-negativity is nontrivial. In addition, it has been demonstrated that the non-negativity constraint reserves high frequency information in the output image [Vogel02]. The total variation regularization with the non-negativity constraint is given by [Krishnan07]:

$$\mathbf{f}_\lambda = \arg\min_{\mathbf{f}\geq 0}\left\{ \| B\mathbf{f} - \mathbf{f}_b \|_{L_2}^2 + \lambda_r \| \sqrt{\mathbf{f}_x^2 + \mathbf{f}_y^2} \|_{L_1} \right\}, \tag{8.6}$$

where $\mathbf{f}$ and $\mathbf{f}_b$ are the original and blurred images in vector format, B represents the blurring filter in vector format, $\mathbf{f}_x$ and $\mathbf{f}_y$ denote the vertical and horizontal image gradients in vector format, and λ_r is the regularization parameter. In the same fashion, the non-negativity constraint can be applied to the lasso regularized deconvolution:

$$\mathbf{f}_\lambda = \arg\min_{\mathbf{f}\geq 0}\left\{ \| B\mathbf{f} - \mathbf{f}_b \|_{L_2}^2 + \lambda_r \| \mathbf{f} \|_{L_1} \right\}. \tag{8.7}$$

8.3.5 Deblurring of outdoor high magnification images

A uniform point spread function (PSF) is used for the deblurring of high magnification images in chapter 6, which works properly for indoor images and outdoor images with an observation distance less than 100m. The recognition rate of the enhanced face images is comparable with that of the image sets collected from a close distance of 1m, as shown in Figure 6.14. However, as the observation distance and system magnification further increase, although the proposed enhancement algorithm is still capable of producing an improved face recognition rate, the performance gap between the high and low magnification data sets remains, as shown in Figure 6.15. A close study of the outdoor high magnification images with an observation distance larger than 100m reveals that such images suffer from nonuniform blurs due to air turbulences. A uniform PSF is unable to accurately describe the actual imaging process. Therefore, multiple PSFs should be used within one image according to the characteristics of the blur. Figure 8.2 illustrates the idea of employing multiple PSFs in one image.

Given an image $f(x, y)$, we first divide it into sub-blocks $f_{i,j}(x, y)$ so that each sub-block undergoes a uniform PSF $h_{i,j}(x, y)$. Afterwards, $h_{i,j}(x, y)$ is estimated and used to deblur the corresponding sub-block. Since there exist abundant PSF estimation and deconvolution algorithms, the key of a successful restoration lies in the proper selection of the block size. A large block size leads to the risk of combining regions with different PSFs. A small block size deteriorates the estimation accuracy of PSFs. Due to the block-

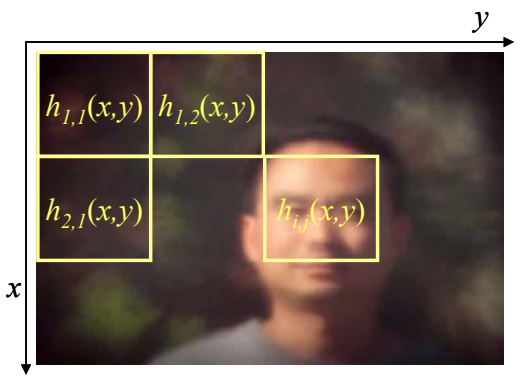


Figure 8.2. The use of multiple PSFs in sub-blocks within one image to compensate for nonuniform blur. The face image is collected from a distance of 300m and with a system magnification of 284×.

based processing, discontinuity may appear at the boundaries of the sub-blocks, which can be alleviated by image blending techniques.

Bibliography

[Abidi05] B. Abidi, Y. Yao, N. Doggaz, and M. Abidi, "Comparison of sharpness measures for the evaluation of image focus", *Workshop on Information Processing and Analysis: Methods and Applications*, Hammamet, Tunisia, Sept. 2005.

[Agarwal07] V. Agarwal, A. V. Gribok, and M. A. Abidi, "Image restoration using L_1 norm penalty function", *Inverse Problems in Science and Engineering*, vol. 15, no. 8, pp. 785-809, Jan. 2007.

[Akaike74] H. Akaike, "A New Look at the Statistical Model Identification", *IEEE Trans. on Automatic Control*, vol. 19, no. 6, pp. 716-723, Dec. 1974.

[Antonini06] G. Antonini, S. Venegas, M. Bierlaire, and J. Thiran, "Behavioral priors for detection and tracking of pedestrians in video sequences", *Int'l Journal of Computer Vision*, vol. 69, no. 2, pp. 159-180, Aug. 2006.

[Batista98] J. Batista, P. Peixoto, and H. Araujo, "Real-time active visual surveillance by integrating peripheral motion detection with foveated tracking", *IEEE Int'l Workshop on Visual Surveillance*, Bomba, India, Jan. 1998, pp. 10-17.

[Batten00] C. F. Batten, "Autofocusing and astigmatism correction in the scanning electron microscope", Mphill thesis, University of Cambridge, 2000.

[Beghdadi89] A. Beghdadi and A. Le Negrate, "Contrast enhancement technique based on local detection of edges", *Computer Vision, Graphics, and Image Processing*, vol. 46, no. 2, May 1989, pp. 162-174.

[Beghdadi04] A. Beghdadi, G. Dauphin, and A. Bouzerdoum, "Image analysis using anisotropic local contrast", *Int'l Symposium on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, Oct. 2004, pp. 506-509.

[Bircheld97] S. Bircheld, "An elliptical head tracker", *Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1997.

[Black01] J. Black and T. Ellis, "Multi camera image tracking", *IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance,* Kauai, Hawaii, Dec. 2001.

[Bouguet00] J. Y. Bouguet, "Pyramidal implementation of the Lucas Kanade feature tracker", Technical report, Intel Corporation, Microprocessor Research Labs, 2000.

[Boult99] T. E. Boult, R. Micheals, X. Gao, P. Lewis, C. Power, W. Yin, and A. Erkan, "Frame-rate omnidirectional surveillance & tracking of camouflaged and occluded targets", *IEEE Int'l Workshop on Visual Surveillance*, Fort Collins, CO, Jun. 1999, pp. 1112-1115.

[Boult03] T. E. Boult, "Geo-spatial active visual surveillance on wireless networks", *Applied Imagery Pattern Recognition Workshop*, Washington, DC, Oct. 2003, pp. 244-249.

[Boult04] T. E. Boult, X. Gao, R. Micheals, and M. Eckmann, "Omni-directional visual surveillance", *Image and Vision Computing*, vol. 22, no. 7, pp. 515-534, Jul. 2004.

[Bretzner00] L. Bretzner and T. Lindeburg, "Qualitative multi-scale feature hierarchies for object tracking", *Journal of Visual Communication and Image Representation*, vol. 11, no. 2, pp. 115-129, Jun. 2000.

[Broaddus05] C. Broaddus, "Statistical Model Selection for Automatic Geometric Camera Calibration", MS thesis, the University of Tennessee, Knoxville, 2005

[Cai99] Q. Cai and J. K. Aggarwal, "Tracking human motion in structured environments using a distributed-camera systems", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1241-1247, Nov. 1999.

[Caviedes02] J. Caviedes and S. Gurbuz, "No-reference sharpness metric based on local edge kurtosis", *IEEE Int'l Conf. on Image Processing*, Rochester, New York, Sept. 2002, pp. 53–56.

[Cielniak03] G. Cielniak, M. Miladinovic, D. Hammarin, L. Göranson, A. Lilienthal, and T. Duckett, "Appearance-based tracking of persons with an omnidirectional vision sensor", *Workshop on Omnidirectional Vision and Camera Networks,* Madison, WI, Jun. 2003.

[Chan98] T. F. Chan and C. K. Wong, "Total variation blind deconvolution", *IEEE Trans. on Image Processing*, vol. 7, no. 3, pp. 370-375, Mar. 1998.

[Chan99] T. F. Chan, G. H. Golub, and P. Mulet, "A nonlinear primal-dual method for total variation-based image restoration", *Journal on Scientific Computing*, vol. 20, no. 6, pp. 1964-1977, 1999.

[Chang01] T. Chang and S. Gong, "Tracking multiple people with a multi-camera system", *IEEE Int'l Conf. on Computer Vision*, Vancouver, Canada, Jul. 2001, pp. 19-26.

[Chen02] X. Chen and J. Yang, "Towards monitoring human activities using an omnidirectional camera", *IEEE Int'l Conf. on Multimodal Interfaces*, Pittsburgh, PA, Oct. 2002, pp. 423-428.

[Choi99] K. Choi, J. Lee, and S. Ko, "New autofocusing technique using the frequency selective weighted median filter for video cameras", *IEEE Trans. on Consumer Electronics*, vol. 45, no. 3, Aug. 1999, pp. 820-827.

[Cindy01] X. Cindy, F. Collange, F. Jurie, and P. Martinet, "Object tracking with a pan-tilt-zoom camera: applications to car driving assistance", *IEEE Conf. on Robotics and Automation*, Seoul, Korea, May 2001, pp. 1653-1658.

[Collins03] R. T. Collins, "Mean-shift blob tracking through scale space", *IEEE Conf. on Computer Vision and Pattern Recognition*, Madison, WI, Jun. 2003, pp. 234-240.

[Cui98] Y. Cui, S. Samarasekera, Q. Huang, and M. Greiffenhagen, "Indoor monitoring via the collaboration between a peripheral sensor and a foveal sensor", *IEEE Workshop on Visual Surveillance*, Bombay, India, Jan. 1998, pp. 2-9.

[Cupillard02] F. Cupillard, F. Bremond, and M. Thonnat, "Group behavior recognition with multiple cameras", *IEEE Workshop on Applications of Computer Vision*, Orlando, FL, Dec. 2002, pp. 177-183.

[Cupillard04] F. Cupillard, A. Avanzi, and F. Bremond, "Video understanding for metro surveillance", *IEEE Int'l Conf. on Networking, Sensing & Control*, Taipei, Taiwan, March 2004, pp. 186-191.

[Dijk02] J. Dijk, M. van Ginkel, R. J. van Asselt, L. J. van Vliet, and P. W. Werbeek, "A new sharpness measure based on Gaussian lines and edges", *Conf. on the Advanced School for Computing and Imaging*, Jun. 2002, pp. 39-43.

[Dockstader01] S. Dockstader and A. M. Tekalp, "Multiple camera fusion for multi-object tracking", *IEEE Int'l Conf. on Computer Vision*, Vancouver, Canada, Jul. 2001, pp.95-102.

[Erdem06] U. M. Erdem and S. Sclaroff, "Automated camera layout to satisfy task-specific and floor plan-specific coverage requirements", *Computer Vision and Image Understanding*, vol. 103, no. 3, pp. 156-169, Sept. 2006.

[Fan03] X. Fan, Q. Zhang, D. Liang, and L. Zhao, "Face image restoration based on statistical prior and image blur measure", *Conf. on Multimedia and Expo*, Baltimore, MD, Jul. 2003, pp. 297-300.

[Faugeras88] O. D. Faugeras and F. Lustman, "Motion and structure from motion in piecewise planar environment", *Int'l Journal on Pattern Recognition and Artificial Intelligence*, vol. 2, no. 3, 1988, pp. 485-508.

[Fayman98] J. F. Fayman, O. Sudarsky, and E. Rivlin, "Zoom tracking", *IEEE Int'l Conf. on Robotics and Automation*, Leuven, Belgium, May 1998.

[Fayman01] J. F. Fayman, O. Sudarsky, E. Rivlin, and M. Rudzsky, "Zoom tracking and its applications", *Machine Vision and Applications,* vol. 13, no. 1, Aug. 2001, pp. 25-37.

[Feris01] R. S. Feris, R. M. Cesar, and V. Kruger, "Efficient real-time face tracking in wavelet subspace", *IEEE Int'l Conf. on Computer Vision*, Vancouver, Canada, Jul. 2001, pp. 113-118.

[Ferrari01] V. Ferrari, T. Tuytelaars, and L. V. Gool, "Real-time affine region tracking and coplanar grouping", *IEEE Conf. on Computer Vision and Pattern Recognition*, Kauai, HI, Dec. 2001, pp. 226-233.

[Fleuret08] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267- 282, Feb. 2008.

[Gonzalez02] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice Hall, Upper Saddle River, New Jersey, 2002

[Grewal01] M. S. Grewal and A. P. Andrews, *Kalman Filtering: Theory and Practice Using MATLAB*, Wiley-Interscience, 2001.

[Griffin05] P. Griffin, "Understanding the face image format standards", *ANSI/NIST workshop*, Gaithersburg, MD, 2005.

[Hager03] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025-1039, Oct. 2003.

[Hatano03] K. Hatano and K. Hashimoto, "Image-based visual servo using zoom mechanism", *SICE Annual Conf.*, Gukui, Japan, Aug. 2003, pp. 2443-2446.

[Hayman03] E. Hayman, T. Thorhallsson, and D. Murray, "Tracking while zooming using affine transfer and multifocal tensors", *Int'l Journal of Computer Vision*, vol. 51, no. 1, pp. 37-62, 2003.

[He03] J. He, R. Zhou, and Z. Hong, "Modified fast climbing search auto-focus algorithm with adaptive step size searching technique for digital camera", *IEEE Trans. on Consumer Electronics*, vol. 49, no. 2, pp. 257-262, May 2003.

[Hoad95] P. Hoad and J. Illingworth, "Automatic control of camera pan, zoom and focus for improving object recognition", *Int'l Conf. on Image Processing and its Applications*, Florence, Italy, Jul. 1995, pp. 291-295.

[Horaud06] R. Horaud, D. Knossow, and M. Michaelis, "Camera cooperation for achieving visual attention", *Machine Vision and Applications*, vol. 16, no. 6, pp. 331-342, Feb. 2006.

[Hu04] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors", *IEEE Trans. on System, Man, and Cybernetics-Part C: Applications and Reviews*, vol. 34, no. 3, pp. 334-343, Aug. 2004.

[Jobson97] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multi-scale retinex for bridging the gap between color images and the human observation of scenes", *IEEE Trans. on Image Processing*, vol. 6, no. 7, pp. 965-976, Jul. 1997.

[Junior02] B. M. Junior and R. Anido, "Object detection with multiple cameras", *Workshop on Motion and Video Computing*, Orlando, FL, Dec. 2002, pp. 187-192.

[Kalka06] N. Kalka, J. Zuo, N. A. Schmid, and B. Cukic, "Image quality assessment for iris biometric", *SPIE Symposium on Defense and Security, Conf. on Human Identification Technology III*, Orlando, FL, Apr. 2006.

[Kang03] J. Kang, I. Cohen, and G. Medioni "Continuous tracking within and across camera streams", *IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, Madison, WI, Jun. 2003, pp. 267-272.

[Kang04] Sangkyu Kang, "Fusion of color and shape for object tracking using PTZ cameras", Technical report, the University of Tennessee, Knoxville, 2004.

[Kannala04] J. Kannala and S. Brandt, "A generic camera calibration method for fish-eye lenses", *Int'l Conf. on Pattern Recognition*, Cambridge, UK, Aug. 2004, pp. 10-13.

[Kehtarnavaz03] N. Kehtarnavaz and H. Oh, "Development and real-time implementation of a rule-based auto-focus algorithm", *Journal of Real-Time Imag*e, vol. 9, pp. 197-203, 2003.

[Kelly95] P. Kelly, A. Katkere, D. Kuramura, S. Moezzi, S. Chatterjee, and R. Jain, "An architecture for multiple perspective interactive video", *ACM Multimedia*, Boston, MA, 1995, pp. 201-212.

[Khan03] S. Khan and M. Shah, "Consistent labeling of tracked objects in multiple cameras with overlapping fields of view", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1355-1360, Oct. 2003.

[Kim03] Y. O. Kim, J. Paik, J. Heo, A. Koschan, B. Abidi, and M. Abidi, "Automatic face region tracking for highly accurate face recognition in unconstrained environments", *IEEE Conf. on Advanced Video and Signal Based Surveillance*, Miami, FL, Jul. 2003, pp. 29-36.

[Kogut01] G. T. Kogut and M. M. Trivedi, "Maintaining the identity of multiple vehicles as they travel through a video network", *IEEE Workshop on Multi-Object Tracking*, Vancouver, Canada, Jul. 2001, pp. 29-34.

[Krishnan07] D. Krishnan, P. Lin, and A. M. Yip, "A primal-dual active-set method for non-negativity constrained total variation deblurring problems", *IEEE Trans. on Image Processing*, vol. 16, no. 11, pp. 2766-2777, Nov. 2007.

[Kristan04] M. Kristan, J. Pers, M. Perse, and S. Kovacic, "A Bayes-spectral-entropy-based measure of camera focus using a discrete cosine transform", *Pattern Recognition Letters*, vol. 27, no. 13, pp.1431-1439, Oct. 2006.

[Krotkov89] E. P. Krotkov, *Active computer vision by cooperative focus and stereo*, New York: Springer-Verlag, 1989.

[Krueger99] V. Krueger, A. Happe, and G. Sommer, "Affine real-time face tracking using a wavelet network", *IEEE Int'l Conf. on Computer Vision*, Corfu, Greece, Sept. 1999, pp. 141-148.

[Krueger00] V. Krueger and G. Sommer, "Affine real-time face tracking using Gabor wavelet networks", *Int'l Conf. on Pattern Recognition*, Barcelona, Spain, Sept. 2000, pp. 127-130.

[Kumar02] R. Kumar, H. S. Sawhney, A. Arpa, S. Samarasekera, M. Aggrawal, S. Hsu, D. Nister, and K. Hanna, "Immersive remote monitoring of urban sites", *SPIE Battlespace Digitization and Network*, Orlando, FL, Apr. 2002, pp. 211-221.

[Kuo02] T. K. Kuo, L. C. Fu, J. H. Jean, P. Y. Chen, and Y. M. Chan, "Zoom-based head tracker in complex environment", *IEEE Int'l Conf. on Control Application*s, Giasgow, UK, Sept. 2002, pp. 725-730.

[Laptev03] I. Laptev and T. Lindeberg, "A distance measure and a feature likelihood map concept for scale-invariant model matching", *Int'l Journal on Computer Vision*, vol. 52, no. 2, pp. 97-120, 2003.

[Lee91] J. Lee, "Analyses of visibility sites on topographic surfaces", *Int'l Journal of Geographical Information Systems*, vol. 5, no. 4, Apr. 1991, pp. 413-429.

[Lee95] J. Lee, K. Kim, Y. Kwon, and H. Kim, "Implementation of a passive automatic focusing algorithm for digital still camera", *IEEE Trans. on Consumer Electronics*, vol. 41, no. 3, pp. 449-454, 1995.

[Lee00] L. Lee, R. Romano, and G. Stein, "Monitoring activities from multiple video streams: establishing a common coordinate frame", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 758-767, Aug. 2000.

[Li02] X. Li, "Blind image quality assessment", *IEEE Int'l Conf. on Imaging Processing*, Rochester, New York, Sept. 2002, pp. 449-452.

[Liao05] Y. Liao and X. Lin, "Blind image restoration with Eigen-face subspace", *IEEE Trans. on Image Processing*, vol. 14, no. 11, pp. 1766-1772, Nov. 2005.

[Lin03] S. Lin, Y. Lin, and C. Su, "An automatic focusing algorithm", *Int'l Journal of Imaging Systems and Technology*, vol. 12, no. 6, pp. 235-238, 2003.

[Linderberg94A] T. Lindeberg, "Scale-space theory: a basic tool for analyzing structures at different scales", *Journal of Applied statistics*, vol. 21, no. 2, pp. 224-270, 1994.

[Lindeberg94B] T. Lindeberg, "Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method of focus-of-attention", *Int'l Journal of Computer Vision*, vol. 11, no. 3, pp. 283-318, 1994.

[Lindeberg98] T. Lindeberg, "Feature detection with automatic scale selection", *Int'l Journal of Computer Vision*, vol. 30, no. 2, pp. 77-116, Aug. 1998.

[Liu03] H. Liu, W. Pi, and H. Zha, "Motion detection for multiple moving target by using an omnidirectional camera", *IEEE Int'l Conf. on Robotics, Intelligent Systems and Signal Processing*, Changsha, China, Sep. 2003.

[Lowe99] D. G. Lowe, "Object recognition from local scale-invariant features", *IEEE Int'l Conf. on Computer Vision*, Kerkyra, Greece, Sept. 1999, pp. 1150-1157.

[Marcenaro01] L. Marcenaro, F. Oberti, and C. S. Regazzoni, "Multiple objects color-based tracking using multiple cameras in complex time-varying outdoor scenes", *IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance*, Kauai, Hawaii, Dec. 2001.

[Mirmehdi97] M. Mirmehdi, P. L. Palmer, and J. Kittler, "Towards optimal zoom for automatic target recognition", *Scandinavian Conf. on Image Analysis*, Lappeenranta, Finland, Jun. 1997, pp. 447-453.

[Mittal04] A. Mittal and L. S. Davis, "Visibility analysis and sensor planning in dynamic environments", *European Conf. on Computer Vision*, Prague, Czech Republic, May 2004, pp. 175-189.

[Morellas03] V. Morellas, I. Pavidis, and P. Tsiamyrtzis, "DETER: detection of events for threat evaluation and recognition", *Machine Vision and Applications*, vol. 13, pp. 29-45, 2003.

[Morita03] S. Morita, K. Yamazawa, and N. Yokoya, "Networked video surveillance using multiple omnidirectional cameras", *IEEE Int'l Symposium On Computational Intelligence in Robotics and Automation*, Kobe, Japan, Jul. 2003, pp. 1245-1250.

[Nguyen01] N. T. Nguyen, S. Venkatesh, G. West, and H. H. Bui, "Hierarchical monitoring of people's behaviors in complex environments using multiple cameras", *Int'l Conf. on Pattern Recognition*, Quebec, Canada, Aug. 2002, pp. 13-16.

[Nummiaro03] K. Nummiaro, E. Koller-Meier, T. Svoboda, D. Roth, and L. van Gool, "Color-based object tracking in multi-camera environments", *Symposium for Pattern Recognition of the DAGM*, pp. 591-599, 2003.

[Ong98] K. H. Ong, "A robust automatic focusing and astigmatism correction method for the scanning electron microscope", Master's thesis, National University of Singapore, 1998.

[Ooi90] K. Ooi, K. Izumi, M. Nozaki, and I. Takeda, "An advanced autofocus system for video camera using quasi condition reasoning", *IEEE Trans. on Consumer Electronics,* vol. 36, no. 3, pp 526-530, Aug. 1990.

[Peixoto98] P. Peixoto, J. Batista, and H. Araujo, "A surveillance system combining peripheral and foveated motion tracking", *Int'l Conf. on Pattern Recognition*, Brisbane, Australia, Aug. 1998, pp. 574-577.

[Peixoto00] P. Peixoto, J. Goncalves, H. Antunes, J. Batista, and H. Araujo, "A surveillance system integrating visual telepresence", *Int'l Conf. on Pattern Recognition*, Barcelona, Spain, Sept. 2000, pp. 98-101.

[Pettre02] J. Pettre, T. Simeon, and J. P. Laumond, "Planning human walk in virtual environments", *IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, Lausanne, Switzerland, Sept. 2002, pp. 3048-3053.

[Phillips02] P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, "Face Recognition Vendor Test 2002, Evaluation Report".

[Poelman97] C. J. Poelman and T. Kanade, "A paraperspective factorization method for shape and motion recovery", *IEEE Trans. on Pattern Recognition and Machine Intelligence*, vol. 19, no. 3, pp. 206-218, Mar. 1997.

[Polesel00] A. Polesel, G. Ramponi, and V. J. Mathews, "Image enhancement via adaptive unsharp masking", *IEEE Trans. on Image Processing*, vol. 9, no. 3, pp. 505-510, Mar. 2000.

[Quan04] L. Quan, Y. Wei, L. Hu, and H. Shum, "Constrained planar motion analysis by decomposition", *Journal of Image and Vision Computing*, vol. 22, pp. 379-389, 2004.

[Quereshi05] F. Z. Quereshi and D. Terzopoulos, "Towards intelligent camera networks: a virtual vision approach", *Int'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Beijing, China, Oct. 2005, pp. 177-184.

[Ramponi98A] G. Ramponi and A. Polesel, "Rational unsharp masking technique", *Journal of Electronic Imaging*, vol. 7, no. 2, pp. 333–338, Apr. 1998.

[Ramponi98B] G. Ramponi, "A cubic unsharp masking technique for contrast enhancement", *Signal Processing*, vol. 67, no. 2, pp. 211–222, Jun. 1998.

[Reid96] I. D. Reid and D. W. Murray, "Active tracking of foveated feature clusters using affine structure", *Int'l Journal of Computer Vision*, vol. 18, no. 1, pp. 41-60, Apr. 1996.

[Roy04] S. D. Roy, S. Chaudhury, and S. Banerjee, "Active recognition through next view planning: a survey", *Pattern Recognition*, vol. 37, no. 3, pp. 429-446, Mar. 2004.

[Roy05] S. D. Roy, S. Chaudhury, and S. Banerjee, "Recognizing large isolated 3-D objects through next view planning using inner camera invariants", *IEEE Trans. on Systems, Man and Cybernetics*, vol. 35, no. 2, pp. 282-292, Apr. 2005.

[Santos97] A. Santos, C. O. De Solorzano, J. J. Vaquero, J. M. Pena, N. Malpica, and F. Del Pozo, "Evaluation of autofocus functions in molecular cytogenetic analysis", *Journal of Microscopy*, vol. 188, pp. 264-272, Dec. 1997.

[Satoh03] Y. Satoh, C. Wang, and Y. Niwa, "Robust object detection for intelligent surveillance systems based on radial reach correlation (RRC)", *IEEE Int'l Conf. on Intelligent Robots and Systems*, Las Vegas, NV, Oct. 2003, pp. 224-229.

[Scotti05] G. Scotti, L. Marcenaro, C. Coelho, F. Selvaggi, and C. S. Regazzoni, "Dual camera intelligent sensor for high definition 360 degrees surveillance", *IEE Vision, Image, and Signal Processing*, vol. 152, no. 2, pp. 250-257, Apr. 2005.

[Seelon96] C. Seelon and R. Bajcsy, "Adaptive correlation tracking of targets with changing scale", GRASP Laboratory Technical Report MS-CIS-96-22, University of Pennsylvania, Department of Computer and Information Science, Jun. 1996.

[Shah04] H. Shah and D. Morrell, "An adaptive zoom algorithm for tracking targets using pan-tilt-zoom cameras", *IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, Montreal, Canada, May 2004, pp. 721-724.

[Shaik00] J. S. Shaik and K. M. Iftekharuddin, "Detection and tracking of rotated and scaled targets using Hilbert wavelet transform", Technical report, 2000.

[Shi94] J. Shi and C. Tomasi, "Good features to track", *IEEE Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, Jun. 1994, pp. 593-600.

[Stainvas00] I. Stainvas and N. Intrator, "Blurred face recognition via a hybrid network architecture", *Int'l Conf. on Pattern Recognition*, Barcelona, Spain, Sept. 2000, pp. 805-808.

[Subbarao92] M. Subbarao and T. Wei, "Depth from defocus and rapid autofocusing: a practical approach", *IEEE Conf. on Computer Vision and Pattern Recognition*, Champaign, IL, Jun. 1992, pp. 773-776.

[Subbarao98] M. Subbarao and J. Tyan, "Selecting the optimal focus measure for autofocusing and depth-from-focus", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 864-870, Aug. 1998.

[Tan94] T. N. Tan, G.D. Sullivan, and K. D. Baker, "Recognizing objects on the ground plane", *Image and Vision Computing*, vol. 12, no. 3, pp. 164-172, Apr. 1994.

[Tarabanis95] K. A. Tarabanis, P. K. Allen, and R. Y. Tsai, "A survey of sensor planning in computer vision", *IEEE Trans. on Robotics and Automation*, vol. 11, no. 1, pp. 86-104, Feb. 1995.

[Tikhonov77] A. N. Tikhonov and V. Y. Arsenin, *Solutions of ill-posed problems*, Jonh Wiley, New York, 1977.

[Tomasi92] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method", *Int'l Journal of Computer Vision*, vol. 9, no. 2, pp. 137-154, 1992.

[Toole05] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abidi, "A video database of moving faces and people", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 812-816, May 2005.

[Tordoff00] B. J. Tordoff and D. W. Murray, "Reactive zoom control while tracking", Technical Report OUEL2228/00, Department of Engineering Science, Oxford University, 2000.

[Tordoff01] B. J. Tordoff and D. W. Murray, "Reactive zoom control while tracking using an affine camera", *British Machine Vision Conf.*, Manchester, UK, Sept. 2001.

[Tordoff03] B. J. Tordoff and D. W. Murray, "Resolution vs. tracking error: zoom as a gain controller", *IEEE Conf. on Computer Vision and Pattern Recognition*, Madison, WI, Jun. 2003, pp. 273-80.

[Tordoff04] B. J. Tordoff and D. W. Murray, "Reactive control of zoom while fixating using perspective and affine cameras", *IEEE Trans. on Pattern Recognition and Machine Intelligence*, vol. 26, no. 1, pp. 98-112, Jan. 2004.

[Utsumi04] A. Utsumi and N. Tetsutani, "Human tracking using multiple-camera-based head appearance modeling", *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, Seoul, Korea, May 2004, pp. 657-662.

[Vogel02] C. R. Vogel, *Computation Methods for Inverse Problems*. Philadelphia, PA: SIAM, 2002.

[Wang04] H. Wang, S. Z. Li, and Y. Wang, "Face Recognition under Varying Lighting Conditions Using Self Quotient Image", *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, Seoul, Korea, May 2004, pp. 819-824.

[Wei01] J. Wei and Z. Li, "On active camera control and camera motion recovery with foveate wavelet transform", *IEEE Trans. on Pattern Recognition and Machine Intelligence*, vol. 23, no. 8, pp. 896-903, Aug. 2001.

[Wei03] Y. Wei and W. Badawy, "A novel zoom invariant video object tracking algorithm", *Canadian Conf. on Electrical and Computer Engineering*, Montreal, Canada, May 2003, pp. 1191-1194.

[Winkler99] S. Winkler and P. Vandergheynst, “Computing isotropic local contrast from oriented pyramid decomposition”, *IEEE Int’l Conf. on Image Processing*, Kyoto, Japan, Oct. 1999, pp. 420-424.

[Wong99] L. M. Wong, C. Dumont, and M. A. Abidi, “Next best view system in a 3D object modeling task”, *IEEE Int’l Symposium on Computational Intelligence in Robotics and Automation*, Monterey, CA, Nov. 1999, pp. 306-311.

[Wong04] C. Wong and M. Kamel, “Comparing viewpoint evaluation functions for model based inspectional coverage”, *Canadian Conf. on Computer and Robot Vision*, London, Ontario, Canada, May. 2004, pp. 17-19.

[Xiong03] Q. Xiong and C. Jaynes, “Mugshot database acquisition in video surveillance networks using incremental auto-clustering quality measures”, *IEEE Conf. on Advanced Video and Signal Based Surveillance*, Miami, FL, Jul. 2003, pp. 191-198.

[Xu05] M. Xu, J. Orwell, L. Lower, and D. Thirde, “Architecture and algorithms for tracking football players with multiple cameras”, *IEE Intelligent Distributed Surveillance Systems*, vol. 152, no. 2, pp. 232-241, Apr. 2005.

[Yagi02] Y. Yagi and M. Yachida, “Omnidirectional sensing for human interaction”, *Workshop on Omnidirectional Vision*, Copenhagen, Denmark, Jun. 2002, pp. 121-127.

[Yamazawa02] K. Yamazawa and N. Yokoya, “Detecting moving objects from omnidirectional dynamic images based on adaptive background subtraction”, *IEEE Int’l Conf. on Image Processing*, Rochester, New York, Sept. 2002, pp. 953-956.

[Yao06A] Y. Yao, B. Abidi, and M. Abidi, “Digital imaging with extreme zoom: system design and image restoration”, *IEEE Conf. on Computer Vision Systems*, New York, Jan. 2006, pp. 52.

[Yao06B] Y. Yao, B. Abidi, and M. Abidi, “3D target scale estimation and motion segmentation for size preserving tracking in PTZ video”, *IEEE Int’l Workshop on Object Tracking and Classification in and beyond the Visible Spectrum*, New York, Jun. 2006, pp. 130.

[Yao06C] Y. Yao, B. Abidi, and M. Abidi, “3D target scale estimation for size preserving in PTZ video tracking”, *IEEE Int’l Conf. on Image Processing*, Atlanta, GA, Oct. 2006, pp. 2817-2820.

[Yao06D] Y. Yao, B. Abidi, N. Kalka, N. Schmid, and M. Abidi, “High magnification and long distance face recognition: database acquisition, evaluation, and enhancement”, *Biometric Consortium Conf.*, Baltimore, MD, Sept. 2006, pp. 1-6.

[Yao06E] Y. Yao, B. Abidi, and M. Abidi, “Fusion of omnidirectional and PTZ cameras for accurate cooperative tracking”, *IEEE Int’l Conf. on Advanced Video and Signal Based Surveillance*, Sydney, Australia, Nov. 2006, pp. 46.

[Yao06F] Y. Yao, B. Abidi, M. Tousek, and M. Abidi, “Auto-focusing in extreme zoom surveillance: a system approach with application to faces”, *IEEE Int'l Symposium on Visual Computing*, Lake Tahoe, NV, Nov. 2006, pp. 401-410.

[Yao07A] Y. Yao, B. Abidi, and M. Abidi, “Extreme zoom surveillance: system design and image restoration”, *Journal of Multimedia*, vol. 2, no. 1, pp. 20-31, Feb. 2007.

[Yao07B] Y. Yao, B. Abidi, and M. Abidi, “Quality assessment and restoration of face images in long Range/High Zoom Video”, *Multi-Sensory multi-model face biometrics for personal identification*, Springer, 2007.

[Yao07C] Y. Yao, B. Abidi, N. Kalka, N. Schmid, and M. Abidi, "Improving long range and high magnification face recognition: database acquisition, evaluation, and enhancement", to appear in *Computer Vision and Image Understanding*.

[Yao07D] Y. Yao, B. Abidi, and M. Abidi, "3D target scale estimation and motion segmentation for size preserving tracking in PTZ video", submitted to *Int'l Journal of Computer Vision*.

[Yao07E] C.-H. Chen, Y. Yao, D. Page, B. Abidi, A. Koschan, and M. Abidi, "Fusion of omnidirectional and PTZ cameras for collaborative surveillance", invited submission to *IEEE Trans. on Circuits and Systems for Video Technology*.

[Yao08A] Y. Yao, C.-H. Chen, B. Abidi, D. Page, A. Koschan, and M. Abidi, "Sensor planning for continuous and automated object tracking with multiple cameras", submitted to *IEEE Int'l Conf. on Computer Vision and Pattern Recognition 2008*.

[Yous06] S. Yous, N. Ukita, and M. Kidode, "Multiple active camera assignment for high fidelity 3D video", *IEEE Int'l Conf. on Computer Vision Systems*, New York, Jan. 2006, pp. 43.

[Zhang99] N. F. Zhang, M. T. Postek, R. D. Larrabee, A. E. Vladar, W. J. Keery, and S. N. Jones, "Image sharpness measurement in scanning electron microscope", *Scanning Part III*, vol. 21, pp. 246-252, 1999.

[Zhang00] Z. Zhang, "A flexible new technique for camera calibration", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no.11, pp. 1330-1334, Nov. 2000.

[Zhu00] Z. Zhu, K. K. Rajaseka, E. Riseman, and A. Hanson, "Panoramic virtual stereo vision of cooperative mobile robots for localizing 3D moving objects", *IEEE Workshop on Omnidirectional Vision*, Hilton Head Island, SC, June 2000, pp. 29-36.

Vita

Yi Yao was born in Nanjing, China, on November 21, 1974, the daughter of Zhongqing Jin and Qi Yao. After graduating in 1992 from Jinlin Middle School, Nanjing, China, she attended Nanjing University of Aeronautics and Astronautics where she received both a Bachelor of Science degree in 1996 and a Master of Science degree in 2000 from the Department of Measurement of Instrumentation. During the summer of 2004, she joined the Imaging, Robotics, and Intelligent Systems Laboratory as a graduate research assistant where she completed her Doctor of Philosophy degree in 2008.