Robust Tracking of Multiple People Using Two Widely Separated Cameras

Chang, Ting-Hsun

QUEEN MARY
AND WESTFIELD COLLEGE
UNIVERSITY OF LONDON

Department of Computer Science

Research Report No. RR-01-04 ISSN 1470-5559 October 2001

Robust Tracking of Multiple People Using Two Widely Separated Cameras

Ting-Hsun Chang

Robust Tracking of Multiple People Using Two Widely Separated Cameras

Ting-Hsun Chang

A thesis submitted for the degree of
Doctor of Philosophy
of the
University of London

Department of Computer Science
Queen Mary, University of London

October 2001

Abstract

The visual analysis of human motion is receiving increasing attention from computer vision researchers who are motivated by its wide spectrum of potential applications, such as man-machine interfaces, video conferencing, and surveillance. An important issue that arises in the automation of many security and surveillance tasks is that of monitoring the movements of people. This thesis addresses the problem of tracking people in an indoor environment.

The automatic tracking system developed in this thesis uses two static, widely separated and un-calibrated cameras to monitor an indoor environment. The tracking task starts with matching subjects' images between successive frames of a single camera. When a camera cannot track the subject well, the tracking information of another camera is used to disambiguate the matching. Thus the system needs to match the subject images across different camera images by establishing feature correspondence.

This thesis applies Bayesian Belief Networks (BBNs) to combine multiple visual modalities for matching subjects' images across camera images. These modalities are based on multi-view geometry, sparse landmarks in the scene and the 2D image appearance of the subject. Gaussian distributions are used to model the feature densities of different modalities for matching subjects across camera images. We also address the problem of lack of colour constancy in a multi-camera system that arises from variations in apparent colour values brought about by different physical processes. To compensate for these appearance variations, the Support Vector Regression (SVR) method is adopted for learning the mapping of visual appearance between two camera images. The benefits of applying BBNs to combine multiple modalities is verified by testing on a large set of sequences and comparison with a naive Bayes method. Experimental results demonstrate that the system can robustly track multiple people and maintain their identities by using two widely separated and un-calibrated cameras cooperatively.

Acknowledgements

The author conducted the research presented in this thesis at the Computer Vision Group in the Department of Computer Science, Queen Mary, University of London. I would like to express my appreciation and gratitude to Professor Shaogang Gong, my thesis supervisor, for his continuous guidance, inspiration and enthusiasm. He manages to strike the perfect balance between providing direction and encouraging independence.

I also would like to thank my colleagues and friends Dr Jamie Sherrah, Yongmin Li, Eng-Jon Ong and Jeffrey Ng for their invaluable advice and useful suggestions on this work. Especially, I really have to thank Eng-Jon Ong for many discussions related to this thesis. The feedback provided by Dr. Sergio A. Velastin, Dr Tim J. Ellis, Professor Yakup Paker, Professor Heather M. Liddell, Professor Guang Li, Dr Richard Howarth, Dr Zhiyuan Luo, Dr Dennis Parkinson and Dr Peter W. McOwan was crucial for the academic progress of the work.

Finally, I am forever indebted to my wife Ya-Hui Cheng for her understanding, endless patience and encouragement, especially taking care of our daughter, Chia-Ling Chang, by herself alone. Without her support, this thesis would not exist. I am also grateful to my parents, without whom none of this work would have been even possible, and to Andrew James Anderson and Keith Anderson for their kind assistance and support.

Contents

List of Figures

List of Tables

Abbreviations

3D	3-Dimension(al)
BBN	Bayesian Belief Network
CPT	Conditional Probability Table(s)
CYM	Cyan, Yellow and Megenta
DAG	Directed Acyclic Graph
FOV	Field Of View
HLS	Hue, Lightness and Saturation
HMM	Hidden Markov Models
HPCA	Hierarchical Principal Component Analysis
HSV	Hue Saturation and Value
JPDAF	Joint Probabilistic Data Association Filter
MCCT	Multiple Camera Cooperative Tracking
MD	Mahalanobis Distance
MHT	Multiple Hypotheses Tracking
MU	Matching Unit
PCA	Principal Component Analysis
RGB	Red Green and Blue
ROI	Region(s) Of Interest
SCT	Single Camera Tracking
SGI	Silicon Graphics
SV	Support Vector
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SVR	Support Vector Regression
VA	Vertical Area
VV	Vertical Volume

Mathematical Notations

arg max	argument that maximises
$\mathcal{X}^2$	chi-square
$\lvert x \rvert$	absolute value of a scalar
$\mathbf{x} \cdot \mathbf{y}$	dot product between vectors $\mathbf{x}$ and $\mathbf{y}$
$\mathbf{x} \times \mathbf{y}$	cross product between vectors $\mathbf{x}$ and $\mathbf{y}$
$E[\cdot]$	statistical expectation
$O(\cdot)$	order of magnitude of
$p(\cdot)$	probability density function
$\lvert \cdot \rvert$	determinant of a matrix
$\lVert \cdot \rVert$	norm of a vector
$\cup$	set union
$\cap$	set intersection
$\in$	element of

Chapter 1

Introduction

The visual analysis of human motion is receiving increasing attention from computer vision researchers. The research is critical to *looking at people* [123, 51], which covers face recognition, gesture recognition, and human tracking motivated by a wide range of applications, such as man-machine interfaces, video conferencing and surveillance. The ability to understand human motion is required if a machine is to interact intelligently and effortlessly with people. Due to the lack of computational power, there were few robust real-time applications during the 1980s [123]. Recently, faster computers have enabled researchers to consider more sophisticated algorithms for real-time human motion analysis. This thesis focuses on one specific problem in this growing field: tracking multiple moving erect people (i.e. walking continuously without extreme motion such as running) in an fixed indoor environment with artificial lighting using two static, widely separated and un-calibrated cameras. An introduction to camera calibration (an process to estimate camera parameters) is given in Appendix A.

1.1 Introduction

Visual tracking has long been studied in computer vision to allow seemingly straightforward human tasks to be carried out by automated systems. For example, many researchers have studied mobile robots that visually locate landmarks, avoid obstacles [142] and navigate within a known environment [143]. Other researchers have focused on perception of vehicles, such as tracking multiple cars in a natural open scene [164] and recognising different types of vehicles [157]. Recently, researchers have begun to address human motion either involving the body parts [83] or the whole body of a moving human without identifying specific parts of the body structure [42, 24].

Most methods for tracking humans use image sequences from a single-camera system which can only cover a limited area from a restricted single viewing angle (camera orientations relative to the object) [3]. Several different approaches have been proposed to relax this limitation, such as use of active cameras [131], wide-angle cameras [68], omni-directional camera systems [111], multiple static cameras [116] or combinations of these different systems [158]. Different approaches have different advantages in terms of tracking moving objects. Tracking using active cameras mounted on pan and tilt platforms enables the system to actively follow moving objects and to provide continuous visual information, whereas the wide angle and omni-directional camera can cover a wider area. The use of several widely separated cameras provides a potential answer to resolving the matching ambiguity by obtaining the scene image from different viewing angles and offers a possible 3-Dimensional (3D) solution [92]. Moreover, a multi-camera system can fuse data from different cameras for a possible interpretation which might not be available from a single camera [15]. Although some researchers have been interested in using multiple active cameras [163], the use of static cameras which do not require ego-motion (i.e. camera motion) estimation can reduce the complexity of a system. This research uses two static widely separated and un-calibrated cameras to track whole bodies of multiple people.

1.2 Problem Domain

Intuitively, increasing the number of cameras might be an effective way to increase the power of a system. However, the complexity of a multi-camera system also introduces two nontrivial issues: (1) how to locate the cameras in a given scene, and (2) how to fuse the data obtained from multiple cameras [129]. The first issue is related to the *sensor planning problem*, namely determining the view-point of different cameras in order to achieve the vision task [162]. One example is the *art gallery problem*, which determines the minimum number of observers (or cameras) necessary to cover a room in an art gallery such that every point in the room is seen by at least one observer [101]. Different methods for locating the cameras influence the functionality and capability that the system can provide. This thesis focuses on the second issue with the aim of fusing multi-camera data in order to make tracking more robust.

To investigate the full potential of using a multi-camera system to track multiple people, this thesis addresses the following problem:

How can we effectively integrate visual modalities, or cues, across different camera images in order to track multiple people using two cameras cooperatively?

By fusing multi-camera data in tracking, a system can reduce visual ambiguity due to additional information provided by different cameras monitoring the same scene. In general, tracking is performed by establishing motion correspondence of objects between consecutive image frames from a single camera. By using multiple cameras to monitor a given scene, the system can utilise the information from different cameras to disambiguate the matching if matching in one camera becomes ambiguous. The system can communicate the tracking information among different cameras in order to maintain visual tracking cooperatively. To communicate tracking information, the system needs to establish the correspondences of different subjects between different cameras. This process of inter-camera subject correspondences can also be used for a system to track and follow people as they move through a large area covered by different cameras [3] (such an extension of monitoring a large area is beyond the scope of this thesis). Although some researchers have proposed solutions to this inter-camera subject correspondences problem, several key issues (discussed in Section 2.2.2) remain that are yet to be properly considered.

1.3 Methodology

Typically, establishing correspondence between two images (also called the *image registration problem* [18]) can be achieved by three general approaches: feature-based, flow-based and iconic-model-based matching [100, 4]. The feature-based approach involves finding a match between the locations of points in two images. The underlying assumption of this method is that the corresponding points can be derived from two images using some low-level operators. The image structural components used by the operators to determine the locations of the points, together with the descriptive attributes, are often referred to as *tokens* [4]. In this thesis, these image structural components and descriptive attributes are called *features*. The flow-based method involves finding a transformation between two images that maps corresponding points between the two images onto one another and uses the brightness constancy assumption to compute the visual motion between two images. The iconic-model method uses correlation templates for matching. It is generally suitable for any type of object, but only when the motion between images is small enough such that the appearances of the corresponding objects in the two images are highly correlated. With two widely separated cameras, a scene is captured from two largely different views. There are significant image variations. Therefore, we consider that the final two approaches are not suitable for our inter-camera correspondence problems. This thesis essentially adopts a feature-based approach to match subjects' images across images from two widely separated cameras.

1.3.1 Feature-Based Matching

In general, feature-based matching includes two steps [100]:

1. extract a set of features from the images (e.g. point, edge, shape, length, orientation, region and colour) to represent the image data, and
2. find correspondences between features in different images which correspond to the same entity in the world, usually called the *correspondence problem.*

Both steps in feature-based matching can be difficult [2]. In the first step, features extracted from different camera images in general correspond to different parts of an object in the 3D world which makes matching difficult. Even if the features corresponding to the same part of the 3D object, they are extracted from images taken from different cameras with different physical processes (e.g. imaging process). The image features from different imaging processes can have different appearances due to different camera parameters (both optical and geometric), viewing geometry (orientation of the object surface normal with respect to the camera and the distance between them) and illumination geometry (orientation of the object surface normal with respect to the illuminant and the distance between them) [159]. The features in different camera images are obtained using different camera coordinates. Therefore, matching features obtained from different camera images would require transformation into a common coordinate system. All these make the correspondence problem extremely difficult to solve and make it the bottleneck of all multi-camera applications [100, 4].

1.3.2 Wide Baseline Stereo Matching

The work presented in this thesis uses two widely separated static cameras to track people. Occlusion causes one of the main difficulties in tracking multiple people consistently over time. Occlusion results from other objects being situated between tracked objects and the camera in such a way that parts or the whole of the target object are not visible. However, by viewing the same scene from two widely-separated viewpoints, occlusions are less likely to occur simultaneously in both camera images. To disambiguate the matching and resolve the occlusion problem, the system needs to establish the correspondences of different subjects between different cameras. This inter-camera correspondence problem is related to the so-called *wide baseline stereo matching* problem where the line between the optical centres of the two cameras (i.e. baseline) is fairly wide, compared to the traditional stereo vision system. Therefore, there are large image variations between the two images to be matched. Traditional correlation-based methods fail to match due to the large differences in images [125]. On the other hand, traditional stereo vision techniques use two cameras for recovering the depth of objects in the scene (distance between the camera and the object), usually with two cameras closely placed [6], leading to small variations between images (e.g. [88]). (A good survey of range vision systems can be found in [69].) The matching techniques employed by conventional stereo are not suitable for wide baseline matching.

Recently, a number of researchers have devised techniques to improve stereo matching, e.g. multiple-baseline stereo methods [85] using several images from different viewing angles simultaneously. Still, individual camera images must generally be close together. In general, stereo with longer baseline yields better depth precision due to wide triangulation, but also increases the likelihood of false matches due to larger image variations. On the other hand, a shorter baseline suffers less precision, but has the advantage of a smaller range of search for the best match [188]. Although wide baseline stereo matching is difficult, it is not only desirable for visual tracking but also for many other applications, such as scene reconstruction with higher depth accuracy. In the context of this thesis, the visual information of an object can be very different in two different camera images, making the matching less reliable. The system therefore requires effective integration of multiple visual modalities in order to make the matching more reliable.

1.4 The Contributions

The work presented in this thesis contributes to the field of visual tracking by developing a multi-camera system to track multiple moving people cooperatively using two widely separated cameras. In particular, the main contributions are:

- Using Bayesian Belief Networks (BBNs) for adaptively combining multiple visual modalities to match subjects from two widely separated cameras.
- The use of multi-view geometry to match subjects across camera images with explicit consideration of the fact that, in an indoor cluttered environment body parts of people can be occluded.
- Using sparse landmarks to infer the spatial relationships of image positions of the corresponding subjects in two camera images. (We define the "two camera images" as two images captured from two static, widely separated and un-calibrated cameras, and use it throughout this thesis).
- Using sparse landmarks to obtain the 3D relative positions of people in a 3D scene with respect to the positions of landmarks and cameras.
- Using the Support Vector Regression (SVR) technique for learning the mapping of the visual information between two cameras in order to compensate for appearance variation and thus make inter-camera matching more reliable. In particular, the lack of colour constancy in a multi-camera system is addressed.

1.5 Thesis Overview

This introduction has given an overview of the research presented in this thesis. The rest of the thesis is organised as follows:

- *Chapter 2* provides a review of the preprocessing, including detection and segmentation of a person or a group of people, and human tracking methods, covering a variety of tracking systems.
- *Chapter 3* gives an overview of the two-camera tracking system, before describing tracking using a single camera involving preprocessing tasks (such as detection, segmentation, feature extraction) followed by tracking based on motion continuity.
- *Chapter 4* describes how to fuse multiple modalities and deal with data uncertainty for matching subjects across two cameras. The BBN is applied to fuse multiple modalities and thus makes matching more reliable. This framework is needed because in practice, visual modalities are unreliable, conflict with each other or only provide partial information. A system has to effectively combine different modalities and cope with such uncertainty.

- *Chapter 5* and *Chapter 6* present details of inter-camera matching techniques. *Chapter 5* focuses on the geometry-based modalities while *Chapter 6* focuses on the recognition-based modalities. Since the system consists of two static cameras, multi-view geometry is adopted to address the inter-camera correspondence problem. Homography (planar projective transformation) and epipolar geometry are used to geometrically constrain the image positions of the corresponding subjects from two widely separated cameras. Since parts of a subject may not be seen in the images in a cluttered indoor environment, the feature-selection problem is explicitly considered in order to apply multi-view geometry effectively. Moreover, scene knowledge, i.e. sparse landmarks, is also employed to aid this matching task.

- *Chapter 6* describes recognition-based modalities to constrain inter-camera subject matching. The appearances of the corresponding subjects in two camera images can be very different, so matching based directly on visual appearance is less reliable. To address this problem, this thesis proposes to learn the mapping of apparent colour and apparent height of subjects' images between the two camera images and use this mapping to estimate the appearance of the corresponding subject across cameras.

- *Chapter 7* demonstrates the results of employing BBN to fuse these geometry-based and recognition-based modalities for matching subjects across different camera images in order to track people using two cameras cooperatively.

- *Chapter 8* concludes the work presented in this thesis, and discusses the limitations of the system and possible further extensions to this research.

Chapter 2

Background Review

This chapter gives an overview of related research for tracking people which includes preprocessing and tracking tasks. The preprocessing task consists of performing detection and segmentation of a person or a group of people in a single camera image and multiple camera images. The tracking task concerns tracking people using a single camera and multiple cameras.

2.1 Preprocessing

This section first discusses preprocessing related to human motion analysis before going into tracking of people. The purpose of preprocessing is to find the Regions Of Interest (ROI) in the image, determining if the regions contain people and segmenting individuals from a group of people or other objects such that the system can perform tracking of the selected regions.

2.1.1 Foreground Detection and Background Modelling

The main purpose of detection is to find selective attentiveness and capture essential visual information in the image. The detected selective attentiveness is then tracked over time based on establishment of correspondence of the selective attentiveness between successive frames of an image sequence. This detection process is critical for a person-tracking system as a system can fail if the subjects are not correctly detected at this stage. Moreover, false alarms can cause a system not to function properly.

This thesis is interested in tracking moving people and their image motion is usually significant enough to be detected [53]. Image motion is one of the most popular cues used to detect people for tracking. In most studies, it is assumed that there is no background motion in the scene, namely the image brightness (intensity) and colour (apparent colour) only changes because of the motion of foreground objects. Thus, the temporal variation in the intensity and colour values reveal the motion of the foreground objects. This temporal change can be detected with a simple pixel-based frame differencing followed by suitable thresholding.

The temporal differencing method is computationally inexpensive but has limitations [156]. For example, it cannot deal with situations involving changes in the background (i.e. objects being introduced or removed from the background), slow moving objects and lighting changes. Some researchers have introduced extensions to the method to alleviate these problems in dynamic scenes. For example, Cheng and Kehtarnavaz [31] used block-wise frame differencing to lower the sensitivity so increasing the system robustness to noise. Olson and Brill [115] used a pre-stored background image without any moving objects as a *reference image* in order to detect any non-background objects. Thus, slow moving objects can be detected. However, without re-initialising the background image, errors in the background image accumulate over time so that the detection is only effective for short-term tracking or in a scene without significant background changes.

To adapt the background image to the current image frames for increasing detection accuracy, a simple method is to average the images over time to obtain a background approximation which is similar to the current background [55]. However, this method can only handle situations where objects move continuously and the background is visible most of the time. It is not effective for scenes with many objects or slow moving objects. Moreover, the rate of adaptation to the changes in the scene is slow and can only have a single predetermined threshold for the whole image frame. Another method involves the use of a linear Wiener filter to predict the intensity value for each pixel level based on recent history [187].

Recently, more complex and robust models for real-time background analysis have been proposed given the increased computational power available. The intensity or colour of each pixel can be modelled as a Gaussian distribution which is adaptive to the changes in the scene using a simple adaptive filter [181] or a Kalman filter [133]. This basic adaptive model does have a pixel-wise automatic threshold using statistical methods, but it cannot handle a background with many non-static objects, e.g. tree branches and leaves in an outdoor scene. In this case, the intensity value varies significantly so that the intensity distribution of each pixel is multi-modal and thus the single Gaussian distribution model usually fails.

In order to cope with the non-unimodal background problem, different approaches have been used. One approach is to classify the image into different components and use a single Gaussian distribution to model each component. Friedman-Russel [50] modelled each pixel value as a weighted combination of three distributions corresponding to three different scene components which are car, shadow and road. Although each pixel is modelled with three different distributions, the scene background (i.e. road) is still assumed to be a unimodal distribution.

Another approach is to consider the background as multi-modal without explicitly considering different scene components. Stauffer and Grimson [156] used a mixture of Gaussian distributions to model the recent history of each pixel. The number of Gaussian distributions is small (in the range of 3 to 5) and is determined by the computational power available. The threshold for background subtraction is determined by the standard deviation of the background model. The pixel outside 2.5 standard deviation is marked as a foreground pixel. This method, using a small number of distributions, cannot handle the case where the background varies with high frequency, since more modes exist and need to be covered in the wide range of variations. On the other hand, a background modelled with a wide distribution can be less sensitive for detecting the foreground object. To cope with this problem, Elgammal et al. [46] used a non-parametric method to model the recent samples of the background. The density estimation with a Gaussian kernel function enables each sample of each pixel in a frame to be considered as a single Gaussian distribution. This method makes the estimation more accurate and quickly adapts to scene changes but can be computationally very expensive.

Colour of human skin can also be used to focus attention on image regions corresponding to the face or hands of a person [53], thus giving an indication of the existence of people. Human face detection has always been an important problem for face, expression and gesture recognition. For example, McKenna et al. [104] built a real-time face detection and tracking system. They have demonstrated that in a Hue-Saturation (HS) colour space (see introduction to HS space in Appendix E), human skin occupies a relatively small cluster and can be used to segment a human head from a complex noisy scene. Yow and Cipolla [186] have also used a colour-based approach to label each pixel according to its similarity to skin colour, and subsequently label each subregion as a face if it contains a large blob of skin-like pixels. However, there are two main limitations in using skin colour for detecting people: (1) there might be skin-like pixels in the background image, and (2) the body parts with skin-colour might not be visible in the image.

Recent studies suggest that data fusion is a promising approach to increasing detection accuracy. Some researchers fused different visual modalities to detect people. For example, motion and colour are used for obtaining better person detection [148, 49]. Other researchers focused on fusing the data from different cameras (i.e. *sensor fusion*). Ivanov et al. [72] used multiple cameras for static background subtraction (no moving background objects). This method used prior knowledge of the pixel-to-pixel correspondence between the background images of two cameras. Since corresponding pixels of the background in two camera images ideally have the same or very similar colour and intensity, the foreground pixels can then be extracted from the corresponding pixels in two camera images with different colour and intensity.

2.1.2 Single Person Detection

Although a degree of selective attentiveness can be obtained by performing perceptual grouping of motion-based and colour-based modalities at the pixel level, such attention in the image is rather crude [53]. It provides a focus of attention without determining whether people are really present in the selected regions of interest. The system needs more knowledge for performing this task. Such knowledge would be used as a model to perform *perceptual search* in the selective regions in order to decide whether the objects in these regions are humans. This process is called "person detection" in this thesis. The main challenge facing a vision-based human detector is the high degree of variability of the appearance of subjects due to articulated motion, partial occlusion and variable clothes texture.

In some applications, person detection might not be that important, such as tracking people in an office environment where there are less likely to be non-humans around. On the other hand, it might become essential in other applications. For example, to monitor an outdoor scene, a surveillance system has to distinguish between humans and non-humans. Cutler et al. [38, 37] applied time-frequency analysis to characterise the periodic motion of people, vehicles and running dogs from airborne video sequences.

They then used the periodicity of motion learnt from the video sequences to classify and track objects in an outdoor scene.

Structural properties of a subject's image can also be used for person detection. A shape-based method was used for this purpose by Cai and Aggarwal [23]. They used multiple cameras to track people in an indoor environment. To distinguish the human images from those of non-humans, they used moment invariants extracted from the silhouettes of subject images as shape features based on Principal Component Analysis (PCA). Lipton et al. [96] extracted and tracked moving targets from a real-time video stream. They first learned the shape properties of image-patches for different objects. In the motion region obtained from temporal differencing, the learnt properties were used to classify the objects into one of three pre-defined categories: human, vehicle or background clutter. Papageorgiou and Poggio [118] reported a pedestrian detection system. They trained a Support Vector Machine (SVM) classifier with local wavelet features derived from a set of training examples of object images (i.e. human and non-human). Their system searched for pedestrians by shifting the detection window over an image and classifying whether the window contained a pedestrian using the SVM classifier. The system has to search the whole image at multiple scales for pedestrians and is computationally extremely expensive, since it does not perform foreground detection to find the regions of interest before detecting people. However, this could also be the advantage of this approach. For example, when the sequence is broken (some image frames missing), is obtained from a moving camera or is after occlusion, a system relying on temporal-differencing does not possess the motion information to detect the foreground whereas this approach can still succeed since it does not need temporal information for detection. They also extended this static image detection technique to a system based on dynamics in images [117]. Instead of learning from a static image pattern, SVM was used to learn the wavelet features of 5 consecutive image frames containing a person.

These kinds of image-based person detection systems have been applied to some practical applications. One example is an automatic airbag deployment system. Krumm and Kirk [90] used visual information to prevent the airbag from operating when the seat of a vehicle is empty or holds an infant. They learnt the images of the passenger seat, taken from a video camera mounted inside the vehicle, based on PCA; and classified the seat as either empty, containing a rear-facing infant seat, or occupied by a person using the learnt eigen-images. One possible limitation of this type of image-based algorithms is the difficulty of generalising the algorithm to different imaging situations with varying pose (i.e. position and orientation), illumination variations and partial occlusion.

Another approach based on multiple sensor modalities has been used to classify a person in a scene [145]. The system performs person detection based on visual and infrared modalities individually and then fuses the results from two sensors. The visual images provide shape information but are difficult to segment due to intensity variations. On the other hand, the thermal image tends to be noisy and the precise shape information is hard to determine. By fusing redundant data of different modalities, the detection accuracy is improved.

2.1.3 Segmentation of a Group of People

All the methods discussed in Section 2.1.2 focused on detection of a single person. Most of the existing systems have assumed that there is only one person in the foreground [37]. However, people tend to interact with each other, and often appear to move as a group. Recently, visual interpretation of groups of people and their motion has begun to attract the attention of researchers in computer vision. However, this visual task is difficult since each individual person in a group is not visually isolated, and usually is partially or totally occluded. For example, Lipton et al. [96] found that their tracking system, based on shape information, tends to mis-classify pedestrians walking together as a vehicle. The difficulty in tracking a group of people lies in the occlusion problem and a system needs to obtain the visual information of each individual in order to track people.

To handle this occlusion problem, different methods have been proposed. Researchers at Carnegie Mellon University [192] developed a stereo-based real-time pedestrian detection system. They used stereo-based segmentation to extract objects from the background and employed a neural network-based method to classify pedestrians in various poses, shapes, sizes, clothing, and occlusion status. The segmentation is performed by finding the discontinuities in the *stereo disparity map*, which usually should occur at object boundaries. The disparity is defined as the difference between the horizontal coordinates of the two corresponding points in the images and is inversely proportional to the depth. They tested the feasibility of their system on the crowded urban streets. This stereo-based segmentation was also used by Darrell et al. [41] to segment a single person from a group of people and background objects in an indoor environment.

In contrast to the stereo-based method, Haritaoglu et al. [57, 58] used silhouette information to segment a group of people into its constituent individuals. They then used a motion model and a correlation-based template matching method to track people. Their system can also classify whether or not a foreground region contains multiple people and can count the number of people in the group by the number of heads in the foreground region.

Other researchers applied the motion-based approach to this problem. Cutler and Davis [37] used the average image size of a person to coarsely segment the foreground to different sub-regions based on the knowledge of the distance between the camera and the people. The homogeneous periodicity of motion is then used to refine the sub-regions in order to count the number of people in the image obtained from the airborne video sequence.

2.2 Tracking People

Having discussed preprocessing in relation to human motion analysis, the work of tracking people is reviewed here. To this end, many criteria could be used to characterise existing methods. For example, they can be based on the types of models used (stick figures, 2D templates, or volumetric models), the dimensionality of the tracking space (2D or 3D), and the sensor configurations (active vs static and single vs multiple). Since this research focuses on tracking people using multiple cameras, the third criterion is adopted to characterise previous work into two classes: (1) single camera based tracking and (2) multiple camera based tracking. Some good reviews based on the first two criteria can be found in [3, 51].

2.2.1 Single Camera Based Tracking

Typically, the methods used for tracking with a single camera build correspondences of the image structures between consecutive frames. This tracking process involves matching pixels, points, lines, contours, and blobs of objects based on their motion, shape, texture, colour and other visual information [4]. As mentioned in Chapter 1, this thesis adopts a featured-based tracking approach which in general needs two steps: feature extraction and feature correspondence. In the first step, image locations satisfying certain well-defined feature characteristics are identified in two consecutive images. This step is very important for tracking because the following matching step is based on the properties of the chosen features. Generally, the criterion for selecting a good feature is its robustness to both noise and appearance change, e.g. size, brightness, contrast. Different types of features have different advantages and weaknesses. Low-level features, such as edge points, are easier to extract but relatively more difficult to track than high-level features, such as contour, region and colour. This is because a large number of candidates for low-level features need to be considered. This can result in matching ambiguity. On the other hand, high-level features are fewer in number and have rich characteristics that can be used for matching, thus increasing matching accuracy. However, the positional ambiguity induced by region-based features can degrade matching. The commonly used geometric features for tracking people are points or blobs. The notion of "blobs" as a representation for image features has a long history in computer vision and has had many different mathematical definitions [113]. In the person tracking context, it is usually obtained from a compact set of pixels that share a visual property that is not shared by its surrounding pixels, e.g. motion, colour, texture. These blobs are often called "motion blobs" or "colour blobs", referring to its visual property.

Detection and analysis of human motion in real time from video imagery have only recently become viable, such as the tracking system Pfinder [181] and W^4 [56, 58]. The general philosophy behind these techniques involves the segmentation of an image, or video stream, into foreground and background regions. After detection, the foreground regions are then classified as human or non-human. The task of tracking people is achieved by matching regions of interest (i.e. subjects' images) between consecutive frames. Most researchers building person tracking systems addressed the problems of detection reliability and tracking despite shadowing and occlusion.

Pfinder ("person finder") [181], the MIT Media Lab's system, is capable of tracking body parts of a person and interpreting their behaviour in real time with a single fixed camera. It has been used as a real-time interface device for many applications, such as video games and virtual reality assuming static background. Technically, the output of Pfinder is the silhouette of a person in the scene obtained by masking out the background. The system first performs temporal differencing by modelling each pixel of background with a single Gaussian model and updating with a simple adaptive filter. The foreground pixels are grouped, based on the spatial and colour similarity. The body parts are then found by employing blob statistics and contour descriptions to roughly indicate the positions of hands, feet, and torso. A feature vector of the blob corresponding to a body part is formulated as (x, y, Y, U, V), consisting of spatial, (x, y), and colour, (Y, U, V), information. After initialisation, the body parts labelling relies on tracking blobs using maximum a posteriori probability approach based on a 2D contour model. However, this system can only track and analyse the motion of a single subject.

W^4 [56, 58], the University of Maryland's system, is designed for real time detection and tracking of multiple people using a single fixed camera. In order to tackle the occlusion problem and maintain the identities of the tracked people, the system uses an appearance template model of the whole body to match the subjects before and after occlusion. However, it was reported that this view-based template does not always apply because the system does not have a mechanism to update the template during occlusion when the appearance of the subject can change significantly [56]. To overcome this problem, the use of a multi-camera system to analyse the occlusion problem was suggested in their previous work [56]. In order to cope with the occlusion problem, the previous system [56] was extended to incorporate silhouette information in order to segment groups of people into individuals [58]. As discussed in Section 2.1.3, the system [58] then tracks the head of each person in the group using correlation-based template matching. However, the head of each individual might not always be visible and the template used to track a person's head can still have the same problem, namely appearance variation during occlusion, at which time the system does not have a mechanism to update the template.

The KidsRoom system [70] at the MIT Media Lab is another real-time person tracking example using a single camera. A notable difference of this system from Pfinder and W^4 is its use of contextual information of the scene. The system tracks multiple children in a play-space by taking advantage of the knowledge of a "closed-world". A closed-world is a space-time domain where the knowledge of all possible objects present in the image sequences is modelled. The system combines four different image features for matching subjects between consecutive frames, i.e. estimated blob size, colour, position and velocity. These modalities are combined assuming independence and with different weights determined by the "closed-world" knowledge. In particular, they considered the correspondence of subjects between consecutive frames globally, i.e. they evaluate the matching for all subjects simultaneously. The reason for matching globally is that local evaluation (i.e. consider the matching of each subject independently) can result in a conflicting match (see Figure 4.2). As in the case of W^4 [56], it was reported that the change of appearance and motion of the subjects during occlusion causes mis-tracking. Moreover, it was also reported that the system has no mechanism for detecting a questionable match. Finally, it was suggested that a change to their system architecture (i.e. a single camera) is required in order to overcome these problems.

Analysis of a dynamic and cluttered scene involves collection of visual evidence extracted from the imaging process, which is almost always subject to uncertainty and incompleteness due to noise, occlusion, and the general ill-posed nature of the inverse-perspective projection [21]. Based on a single modality of visual information to track multiple people, it can be less reliable. For example, Rosales and Sclaroff [135] used an extended Kalman filter to track multiple moving people and used the predicted positions to resolve the occlusion problem. It was reported that the use of the motion modality alone is insufficient for handling certain tracking scenarios, such as people changing direction when they meet and the occlusion is present in the image. This is because the motion of a walking human is difficult to model due to the potential instability of human motion. Another relevant example is tracking based on the colour modality alone. Khan and Shah [87] used a mixture of Gaussian distributions to model the colour of whole blob and tracked people based on the colour. However, their system cannot reliably handle situations where people's clothes have the same or similar colour.

Although tracking can be made more robust by combining multiple visual modalities, difficulties still remain due to occlusions or a cluttered background, as reported in the KidsRoom system [70]. This difficulty is mainly due to no mechanism available for updating the subject appearances during occlusion. Some researchers have tried to overcome this limitation by using different types of cameras or a careful choice of camera location. Naya [111] used an omni-directional camera composed of a parabolic mirror and a video camera to obtain a panoramic image in order to track multiple people in a room. There are, however, two limitations to this approach. First, the low resolution of the image may not provide enough visual information for subsequent visual tasks, e.g. face recognition. Second, the visual information is still limited to a single viewpoint. Other researchers attempted to increase the mobility of the system by using an active camera to track people [113]. In general, this type of architecture is designed to track a single person by actively controlling the camera. Again, the visual information of the system is obtained from only one viewpoint. Rossi and Bozzoli [138] mounted a camera vertically to track and count people from the ceiling. Occlusion of multiple subjects was avoided due to the viewing angle of the camera. The system tracks people based on position estimation. However, the monitoring area without occlusion in this case is limited to the area not far from the camera. In short, a single-camera system has limited access to useful visual information. The occlusion problem largely remains difficult to resolve.

2.2.2 Multiple Camera Based Tracking

Only recently have researchers begun to use multiple cameras for human motion analysis. There are different reasons to use a multi-camera system, such as increasing the monitored area and/or obtaining visual information from different viewpoints. This thesis uses multiple cameras because simultaneous occlusions in both cameras are less likely given occlusion in one camera. Figure 2.1 shows an example of a camera setup where two cameras have an overlapping FOV and a person, shown as an ellipse, appears in the overlapping FOV. When another person is in the shaded areas A_a and A_b in the diagram on the left, one or other person will be occluded from the point of view of the left camera. Given occlusion in the left camera, however, occlusion will only occurs in the right camera image, if the second person appears in the areas A_c or A_d (see the diagram on the right). Assuming the probabilities that people appear at different places in the overlapping FOV are equal, the probability of occlusion in the right camera given occlusion in the left is equal to $\frac{(A_c+A_d)}{(A_a+A_b)}$. From these diagram, it can be seen that the area $(A_c + A_d)$ are far smaller than $(A_a + A_b)$. Thus, the probability of simultaneous occlusion in both cameras is small given occlusion in one camera. In the following, the multi-camera systems containing active cameras are first reviewed before we look into multiple static-camera systems.

Multi-Camera Systems Containing Active Cameras

The first option is to use two active cameras, often referred to as *vergence stereo*, to track an object and keep the optical axes intersecting at the same surface point of the object. This process of actively controlling camera motion so that a 3-D point in the world is constantly at the same point on the image plane is called *camera fixation* [131]. This type of binocular vision system is a modification of the conventional stereo approach [6]. The advantage of this type of configuration is that it can obtain the depth information and also obtain the high resolution region of image points of interest. However, the system needs some mechanisms for gaze control in order to aim both cameras at a given

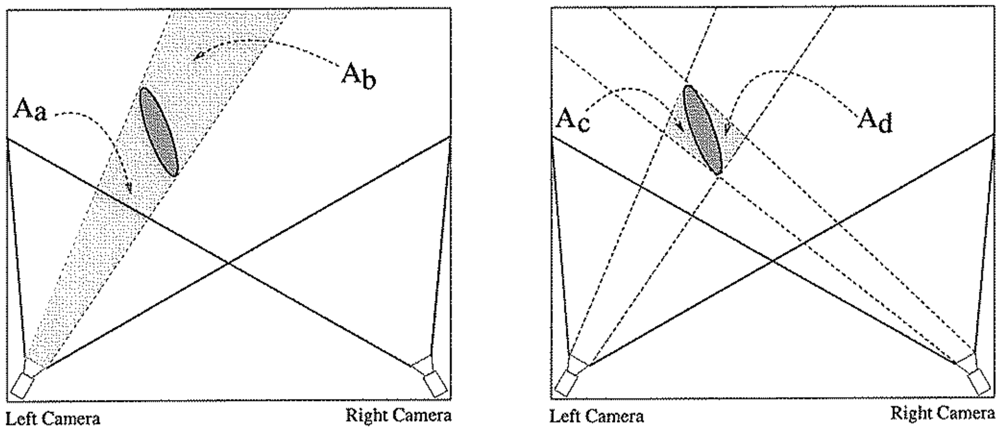


Figure 2.1: *Simultaneous occlusions in both cameras are less likely given occlusion in one camera.*

point of the object. The vergence stereo has applications in autonomous robots tracking and navigation [163].

The other system option is the hierarchical structure where different cameras are responsible for different stages or different parts of the system operation. A system uses the data from some cameras to guide the operation of other cameras. For example, a system can use some cameras for searching for targets in order to guide some others for tracking these targets, called *searching* and *tracking* cameras respectively. The searching cameras provide the system with only approximate and global information concerning the environment which is then used to guide the tracking cameras to focus on a narrow area of interest in the environment. This operation of sensor fusion is often referred to as *guiding* or *cuing* [97]. The same mechanism is also used in other fields, e.g. military surveillance system. Huang et al. [68] reported a tracking system with two calibrated cameras, consisting of a wide-angle camera and a narrow-angle camera. The fixed wide-angle camera is used to monitor a larger area in the scene and guide the narrow-angle active camera to track a moving person and provide a high-resolution image of the person's face. This strategy was also adopted by Stillman et al. [158] and Peixoto et al. [121] to track multiple people in an indoor environment, and by Kenneth and Dawson-Howe [43] to track pedestrians on the street. Stillman et al. [158] used two fixed searching cameras and two active tracking cameras to track people in a room. Peixoto et al. [121] used one wide-angle static camera to guide a vergence stereo. Kenneth and Dawson-Howe [43] used a wide angle camera to guide an active camera.

Multi-Camera Systems Containing Static Cameras

Another category of multi-camera system consists of multiple static cameras, which independently provide information about the same scene and are in competition as to which will be believed by the system. The sensors of this type of configuration are often referred to as *competitive sensors* [15, 45]. The system configuration adopted in this thesis belongs to this category. This type of system can interpret the scene more reliably due to redundant information of the same scene which enables the system to have the potential to remove uncertainty in the data. To integrate redundant information, the system needs to determine what information from different cameras corresponds to the same entity in the scene. This is the correspondence problem, also known as *sensor registration* in the sensor fusion literature [98]. This inter-camera correspondence problem is similar to the *data association* problem [9] in target-tracking which associates the data between consecutive image frames from a single sensor over time. The difference is the former needs to associate the data obtained from different sensors which can be more difficult than the latter due to the data being obtained from different sensor coordinates and different physical processes.

One of the earliest systems for tracking multiple people using multiple cameras was reported by Rao and Durrant-Whyte [128, 129]. They used four fixed cameras to monitor a room and used Kalman filters to track multiple people on the ground plane. The subject image is first obtained by background subtraction. The image position of the subject's lowest point is transformed into floor coordinates which is fed to the tracker. The tracking is then performed on the ground plane. The correspondence of subjects between consecutive frames of a camera is determined by searching for the subject with the closest distance between the observed position and the predicted position from a Kalman filter. This matching criterion, known as the *nearest neighbour approach* [9], is widely used for addressing data association and motion correspondence problems. To match subjects across cameras, the system used the transformed position of the subject's lowest point on the ground floor based on the same method, i.e. nearest neighbour approach. This constraint is often referred to as the *ground plane constraint* in the literature [81, 92], which assumes that the 3D position of an object lies on the ground plane and each 2D image point of the subject corresponds to a unique 3D point. This ground plane constraint is widely used for inter-camera matching. For example, Jones and Giaccone [81] used this constraint for matching subjects between cameras to track people in a car park, and Kelly et al. [86] used the same constraint for tracking people on a campus courtyard with four cameras. Other researchers extended this constraint to further incorporate a world model (i.e. model of the environments) to track multiple people in a building with distributed cameras [140]. However, the assumption of a ground plane constraint might not hold when a person's feet are not in the camera images. This situation is important especially for tracking people in an indoor environment with clutter. Moreover, these systems can have limitations since only position information is used for matching subjects between images from different cameras.

Instead of transforming the image positions to a common coordinate system (e.g. ground plane), some researchers applied multi-view geometry to the inter-camera correspondence problem. Multi-view geometry provides the geometric relations that exist between the object images in different views. This multi-view geometry-based approach has progressed remarkably in the last decade [59]. Some multi-view geometry, e.g. epipolar geometry and homography (planar projective transformation), can be recovered without calibration and only needs the knowledge of a set of sparse corresponding points in different views. This simple requirement makes multi-view geometry a simpler algorithm than traditional calibration-based methods [59]. More details relating to multiple-view geometry are discussed in Chapter 5.

To apply multiple-view geometry to match people in different camera images, different researchers used different features. Meyer et al. [105] applied the ground plane constraint and used the homography to map the lowest point of the subject image from one camera image to the other camera image in order to keep tracking people as they walk along the FOVs of different cameras in an outdoor scene. Instead of using the ground plane constraint and subjects' lowest point for applying homography, Lee and Stein [92] used silhouette centroids of objects and assumed these points all lie on a virtual plane about one meter above the ground. The reason for using this feature point is to avoid segmentation errors, caused by for example the shadow effect, seen when using the lowest part of the subject image. They also used the knowledge of the intrinsic parameters and the homography between two images of two cameras to map the virtual plane in each camera image into a single overhead image for global activity understanding. However, they noted that the centroids in different camera images could correspond to different 3D points and these points might not lie on the same plane in the world.

Since homography only applies to 3D coplanar points in the scene, the subject's feature points used for matching across camera images must lie on the same plane. However, this common plane for applying homography might not always be available in the scene. Some researchers use other types of geometry regarding two perspective views.

Cai and Aggarwal [22, 23] used epipolar geometry for matching people across cameras in an indoor environment where most of the ground is not visible. Epipolar geometry can simplify the correspondence problem by reducing the search space from a 2D image plane to a 1D epipolar line [182]. The inter-camera correspondence of a 2D image point in one camera image can be performed by searching along its corresponding epipolar line in the other camera image. To apply the epipolar geometry, they used multiple points extracted from the medial axis of the subject's upper body as the feature points. Their system can also perform camera switching in order to keep tracking people as they walk through a large indoor environment covered by distributed cameras. This switching consists of two steps: predicting when people leave the FOV and selecting a camera for tracking people. For the same tracking scenario, other researchers [75] used the geometric knowledge of the boundary of FOV to predict when people move out the FOV of a camera. They also use the knowledge of the FOV boundary of one camera in the other camera images to disambiguate the matching when people walk through different FOVs.

Other researchers used modalities based on recognition to match subjects across cameras. For example, Collins et al. [34] used a normalised colour histogram of objects' images and their 3D trajectories on the ground plane to match vehicles and people across camera images. They use a colour histogram generated from object images to search for the closest match after occlusion in order to maintain tracking. Other researchers [116, 158] used colour modality for inter-camera matching to track people with multiple cameras in both outdoor and indoor environments. However, all these methods directly applied apparent colour of a subject's image from one camera to match to its corresponding subject's image from the other camera. Not transforming the colour obtained from one camera image to a suitable value for the other camera image causes unreliable matching. This is because the apparent colour of an object can vary significantly when captured by different cameras. In general, the apparent colour of an object depends on the illuminants, the reflectance of the object, illumination geometry, viewing geometry and camera parameters [20]. Although human vision has the ability to adapt to changes in colours, known as *colour constancy*, machine vision has yet to find a reliable means to compensate for this shift in order to recognise objects by colour. Therefore, the colour constancy problem of a multi-camera system needs to be addressed.

2.3 Summary

This thesis focuses on developing a methodology to track multiple people in the overlapping FOVs of two widely separated and un-calibrated cameras. Based on our discussions of existing techniques, it is clear that one of the main difficulties in tracking multiple people is caused by the lack of continuous visual information for each individual under occlusion or when people are moving as a group. Rosales and Sclaroff [135] reported that the extended Kalman filter fails to correctly track multiple moving people when people change walking direction during occlusion. The W^4 [56, 58] and the KidsRoom [70] system, systems discussed in Section 2.2.1, also reported that significant appearance variations during occlusion results in the degradation of tracking accuracy due to the lack of an update mechanism for the state of the subjects' appearance and motion.

One possible solution for this occlusion problem is to segment merging blobs containing multiple people into individuals as discussed in Section 2.1.3. However, the assumption of this approach does not hold when the segmentation fails or people are not visible from a certain viewpoint. Another possible solution to this occlusion problem is to exploit multiple cameras. This approach is more promising because of the existence of redundant information from multiple viewpoints. This multi-camera method assumes that the system always has an unambiguous image of each individual in some cameras when occlusions are present in others. Therefore, the multiple static cameras need to be widely separated with significantly different viewing angles such that the occlusions are unlikely to be present in all cameras at the same time.

The idea of using more cameras to disambiguate the matching ambiguity is not new. For example the trinocular stereo [54] uses a third camera to solve the matching ambiguity between two camera images in the stereo matching problem. Ng et al. [112] also extended this technique to *N-Ocular Stereo* in which four omni-directional cameras were applied to track multiple people. However, the underlying problem of using multiple widely separated cameras to cooperatively track people, i.e. wide baseline stereo matching, is not addressed. Without knowing "who is who" between different cameras, the system cannot disambiguate the matching in each single camera.

To match subjects across cameras, some researchers used the multi-view geometry (e.g. homography and epipolar geometry) without considering the feature selection problem, namely when the lower part of the subjects are not visible in the image. This thesis will consider this problem, and investigate what other geometry-based modalities can also be exploited. Other researchers used recognition-based modalities (such as colour) without considering the variations in the subject appearances between different cameras. This research will address the colour constancy problem in the multi-camera system and explore what other image patterns of a subject can be used. Moreover, we also address the problem reported in the KidsRoom system [70], that the system has no mechanism for detecting a questionable match. Detecting this matching ambiguity will be considered for both cases: (1) matching between consecutive images obtained from a single camera and (2) matching across different cameras' images. The methods used to detect the ambiguity in these two matching tasks are given in Section 4.4.3.

Most existing multi-camera systems either use one single visual modality [128, 129, 68, 86, 81, 92, 105, 75, 116, 158] assume all modalities are independent [22, 23], or do not explain how to fuse different modalities [140, 115]. As discussed in Chapter 1, the inter-camera correspondence problem between widely separated cameras is difficult, and a system should use multiple visual modalities and consider the correlation (or dependency) between different modalities to make matching more reliable. Moreover, since the data is always uncertain and different modalities might conflict with each other, a framework is required to fuse all modalities to improve the inter-camera subject correspondences. We use a BBN (Bayesian Belief Network, see Chapter 4) for fusing multiple modalities, capturing the correlation between different modalities and handling data uncertainty. This BBN is also used to obtain the global consistency in the correspondence problem noted in the KidsRoom tracking system [70] (as discussed in Section 2.2.1), where all modalities are assumed to be independent.

Chapter 3

System Overview and Single Camera Tracking

3.1 The System Architecture

Considering the problem of tracking multiple people in an indoor office environment, the background objects and lighting conditions are relatively stable. Since a single static camera has a limited viewing angle, two widely separated cameras (see Figure 3.1) positioned at the neighbouring corners of the room are used, so that the image variations between two camera images are large and occlusions are less likely to happen in both cameras simultaneously.

Figure 3.1 shows an example of such a setup where the cameras have an overlapping FOV and the system aims to track people in this overlapping area. The room is about 6 meters long 4 meters wide and 2.8 meters high and both cameras are located 2.45 meters high at the neighbouring corners. In the overlapping area (3 by 4 meters), the maximum number of people can be imaged in the room without occlusion in a camera image is about 4 people.

The cameras are static and un-calibrated so that their parameters and relative positions are fixed but unknown. This is because calibration information may not be practically available in some circumstances [59]. To relax reliance on the calibration process, the stereo algorithm is not used in our system, though using depth information can make tracking more reliable than 2D information alone. Moreover, if people are close to each other in the scene, tracking reliant on depth alone might result in matching ambiguity. To simplify the control problem, the image streams from the two cameras

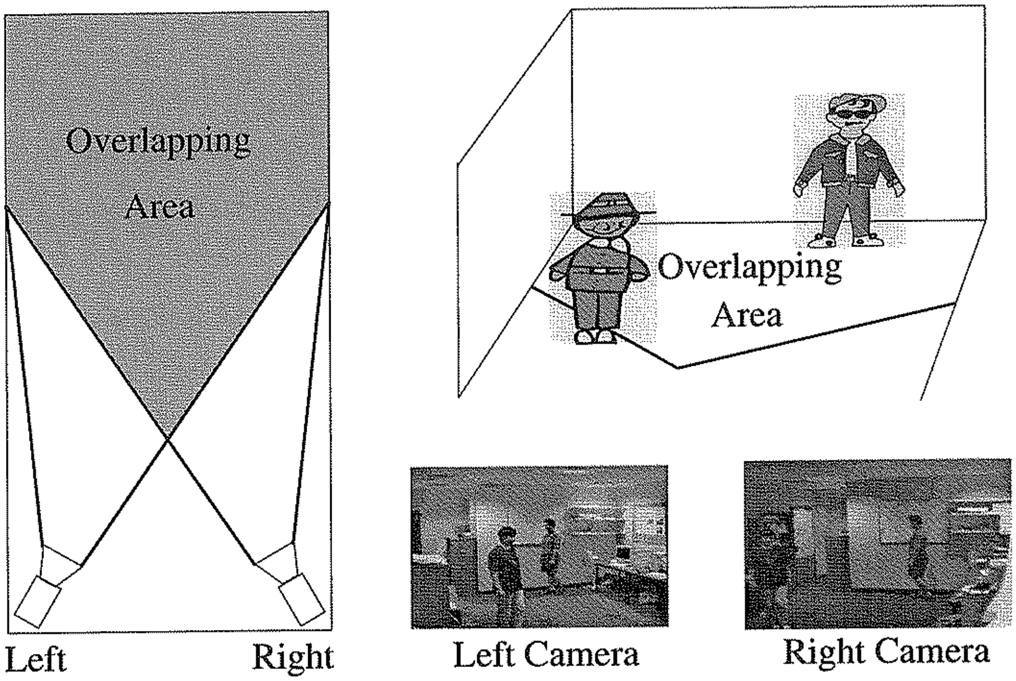


Figure 3.1: *An example of a two-camera tracking system.*

are not synchronised. Thus, the two images from two cameras are not captured at the same time instant. Note that neither of the two cameras can cover the whole room. As a consequence, the lower body parts of subjects might not always be visible in the camera images so that the ground plane constraint cannot always apply (see Figure 3.1). Moreover, occlusion can also often be present in a cluttered indoor environment. This occlusion problem can make the tracking task difficult due to incomplete information being available in the image.

Given such a camera setup, our goal is to segment the images of moving subjects from the background and then to track moving people in the overlapping area over time from an image sequence pair of two monocular cameras. (We define the "an image sequence pair" as two sequences captured from two static, widely separated and un-calibrated cameras, and use it throughout this thesis). The system first tracks the subjects in each camera based on its own visual information. To track people using two cameras cooperatively, the system assigns a label to each newly detected subject. This label is referred to as the *identity* of a subject. The system aims to track people with the assigned identities over time using two cameras cooperatively. If a newly detected subject in an image of a camera I_i has already been tracked and assigned an identity in the other camera image I_j, the system then passes the identity to this subject in I_i by matching subjects across camera images. This process of matching subjects across camera images can also be used to regain the identity of a subject from other cameras. Moreover, it can be used to check whether different subjects with the same identity in different cameras correspond to the same person. Thus, our system has two major tracking modes:

1. Single Camera Tracking (SCT) which matches the subject images between successive image frames of a camera over time, and
2. Multiple Camera Cooperative Tracking (MCCT) which matches subject images across cameras to establish correspondence of subjects between two camera images.

The relationship between these tracking modes is shown in Figure 3.2. The system performs SCT to track people as long as they are in the FOV. The MCCT mode matches subjects' images across cameras to pass the identities between cameras.

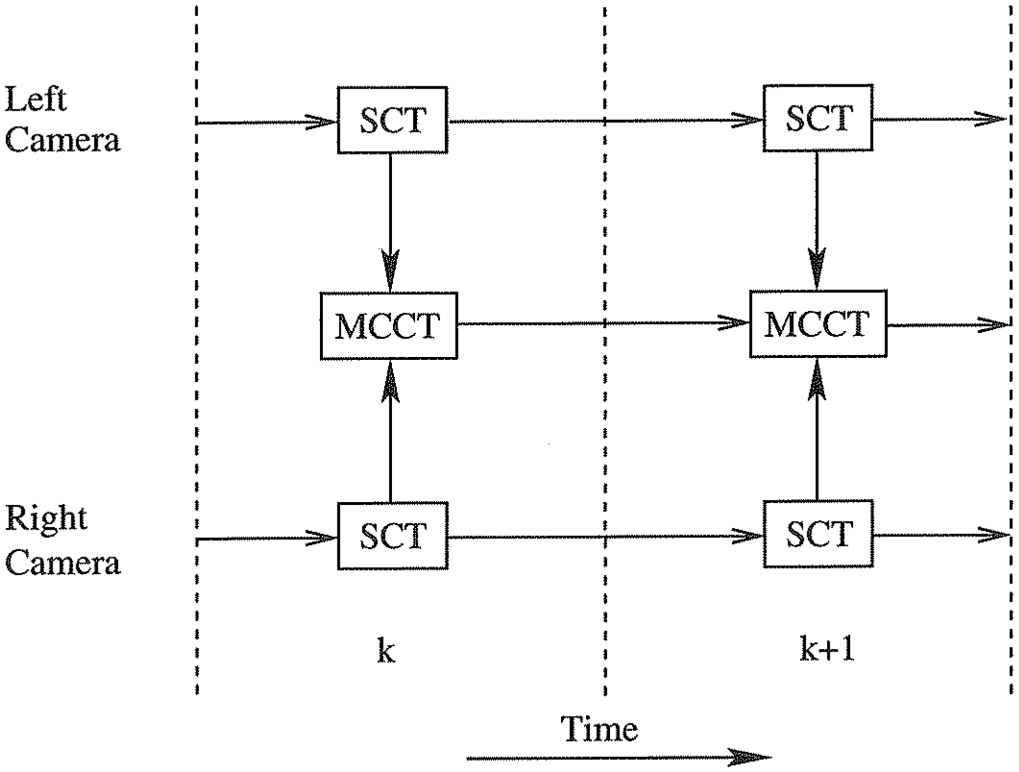


Figure 3.2: *The relationships between tracking modes.*

The tracking scenarios are shown in Figure 3.3. The goal is to track people with identities over time using two cameras cooperatively. The unshaded box represents the case when a camera captures and tracks people with identities. The unshaded box with a question mark represents the case when a camera captures people without identities, e.g. subjects re-appear in a camera image after occlusion and the subject identities are not maintained. The shaded box indicates the case when the camera image is ambiguous, e.g. occlusion is present in the camera image or the camera is subject to tracking failure. Figure 3.3 shows two tracking cases where MCCT is needed. In case 1, people are initially in the FOV of the left camera but not in the right camera. The system assigns identities to each individual and passes these identities to the subjects in the right camera once people enter the FOV of the right camera. In case 2, the system passes the identities maintained in the left camera to the right camera in order to resolve the matching ambiguity when the system has lost the subject identities in the right camera image.

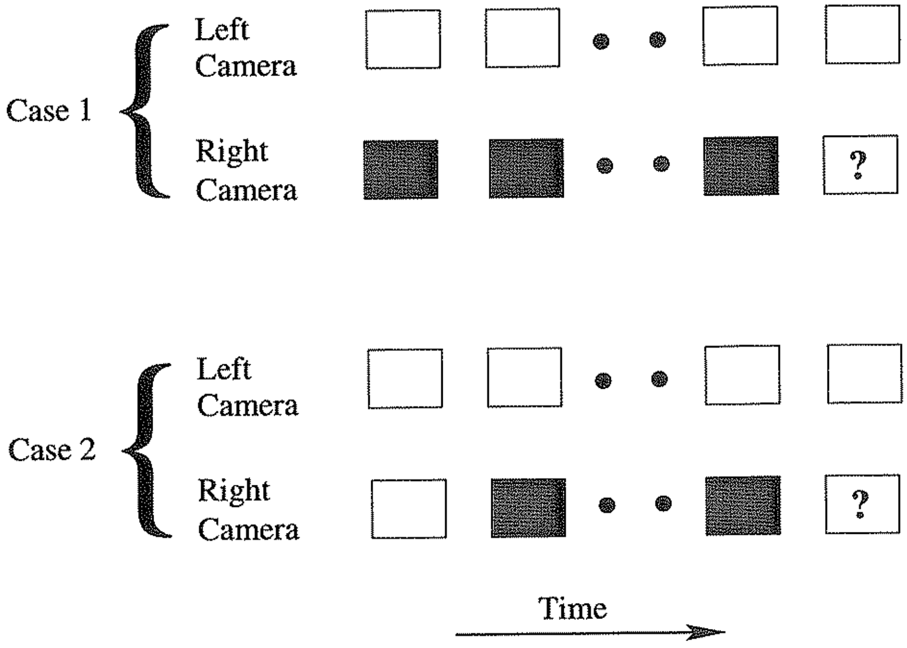


Figure 3.3: *An illustration of the proposed tracking problem for MCCT.*

3.2 Single Camera Tracking

As discussed above, the system has two different tracking modes: Single Camera Tracking (SCT) and Multiple Camera Cooperative Tracking (MCCT). The rest of this chapter describes the SCT in detail and the MCCT is presented in the following chapters. Tracking with a single static camera includes two major steps: preprocessing and matching of subjects between successive frames (Figure 3.4). Two stages of preprocessing are performed: (1) segmenting the moving subjects from a still background (see Figures 3.5 and 3.6) and (2) extracting feature points from the segmented subjects' images (see Figure 3.7). After preprocessing, the system establishes the feature correspondences between consecutive image frames for tracking people based on Kalman filtering.

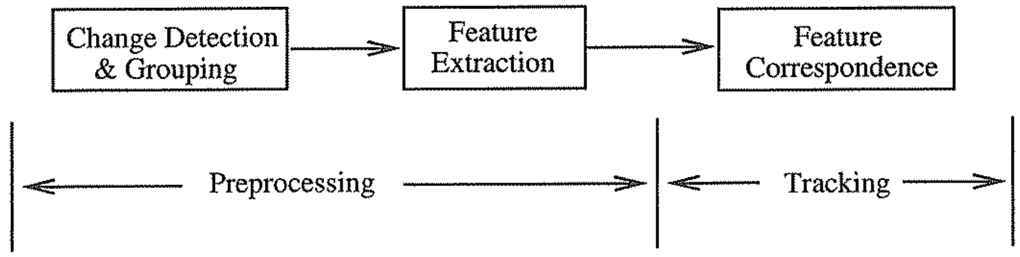


Figure 3.4: *Block diagram of tracking using a single camera.*

3.2.1 Preprocessing

Change Detection and Grouping

To detect foreground objects, we take advantage of the fact that the camera is stationary. Therefore, moving objects can be segmented using a simple frame differencing method assuming background objects and lighting condition are largely stable. The intensity of the current image $I(x, y, k)$ is subtracted from the pre-stored background image $B(x, y)$. If the intensity change at (x, y), $(|I(x, y, k) - B(x, y)|)$, is above a predetermined threshold, $\mathcal{T}$, it yields a foreground pixel. This *background subtraction* is computationally inexpensive. However, it is sensitive to noise in the imaging process and can be degraded by the effect of shadows and reflections so that a suitable threshold is hard to find [53, 136]. The value of the threshold used in the experiment is set manually. However, the limitation of this chosen threshold is that it typically only works well for the environment where the experiments are conducted in a certain lighting condition. For example, when the lighting condition changes the threshold might need to be re-adjusted. The morphological operations, two times erosion followed by one time dilation (both with 3×3 structuring element) [155], are performed to alleviate the noise problem. Figure 3.5 shows an example of applying background subtraction to detect the foreground pixels. Each row contains two images from the left and the right cameras. The top row shows the pre-recorded background images. The middle row shows two

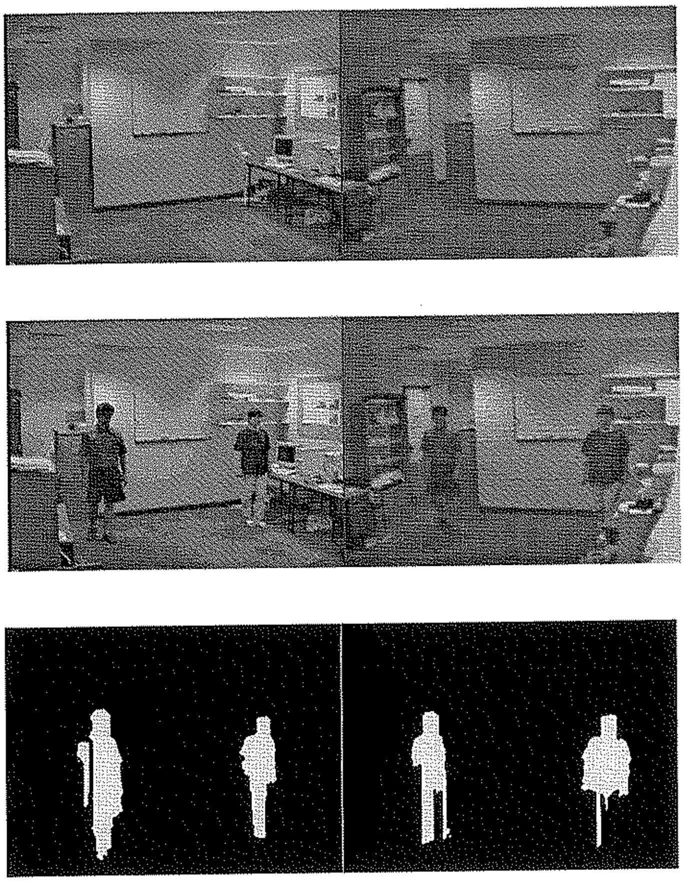


Figure 3.5: *Moving object detection.*

people walking in the scene and are observed in two camera images. The bottom row shows the binary images after change detection by differencing, thresholding and noise reduction. Note that the lower parts of the subjects may not be segmented well due to shadow, occlusion by other objects in the scene and the improper threshold used for differencing. The binary image (after noise reduction) containing the detected pixels is called the *foreground image*, where the intensity value of each pixel, $F(x, y, k)$, is defined as:

$$F(x,y,k) = \begin{cases} 1 \quad , & |I(x,y,k) - B(x,y)| \geq \mathcal{T}. \\ 0 \quad , & |I(x,y,k) - B(x,y)| < \mathcal{T}. \end{cases} \tag{3.1}$$

Once the images of the non-background objects are separated from the background image, the next step is to locate the images of foreground objects within different bounding boxes to provide a focus of attention for further processing. The technique of *equivalence classes* [124] is applied to group the foreground pixels into different blobs as follows:

1. The foreground image (400×300) is first divided into a coarse grid of 20×15 equal size bins (i.e. small blocks) in order to reduce sensitivity to noise and computational expense. Each bin is labelled with the number of foreground pixels contained within it.

2. These labels (i.e. number of foreground pixels) are then thresholded to obtain a set of bins considered to be regions of foreground pixels.

3. These regions are then grouped into equivalence classes which are the detected foreground objects.

4. Each detected blob is then circumscribed by a bounding box.

Figure 3.6 shows the bounding boxes obtained by grouping the foreground pixels in the binary images of Figure 3.5.

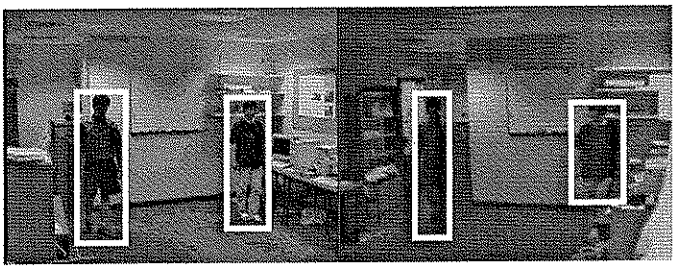


Figure 3.6: *Two images from a sequence pair captured by two widely separately cameras with bounding boxes for the moving objects.*

Feature Extraction

Given the blobs that have been detected in the image, the next step is to extract features from the subject image for establishing subject correspondences between consecutive frames. One possible method is to measure the optical flow and use it to guide the matching process, such as in [183]. However, this method assumes constant brightness between images. Thus it may not be appropriate for tracking a non-rigid human body which can undergo significant appearance changes [135]. Alternatively, feature-based method is used for tracking subjects. To match subject images, different features can be selected, such as colour [87], a template of the subject image [57, 56], a feature point of the bounding box [135], or feature points of the subject image [24]. Different features have different advantages and weaknesses. For example, colour is robust with respect to common geometric distortions (e.g. rotation, translation, cropping, scaling) [14], but cannot reliably handle the subjects' images of similar colours for establishing subject correspondences. Tracking can be made more robust by fusing different feature modalities. Since this thesis focuses on matching subjects between two cameras, only one feature modality (i.e. the highest point of the subject image) is used for matching subjects in SCT and the colour modality is not used. Compared to other points, such as the centroid or the lowest point, the highest point usually corresponds to the same 3D world point and is more robust to the effect of shadow or reflection. This advantage enables the system to track a subject more smoothly and robustly. An example of the extraction of the highest points from two subject images is shown in Figure 3.7. Figure 3.8 shows an example where the extracted lowest points in two camera images do not correspond to the same point of a person in the 3D world. This is because the lower part of the person is out of view of the left camera. Figure 3.9 shows an example where the lowest points in two consecutive frames captured from the left camera do not correspond to the same point of the person in 3D world. This is because the lower part of the person is occluded by a table in the left camera view.

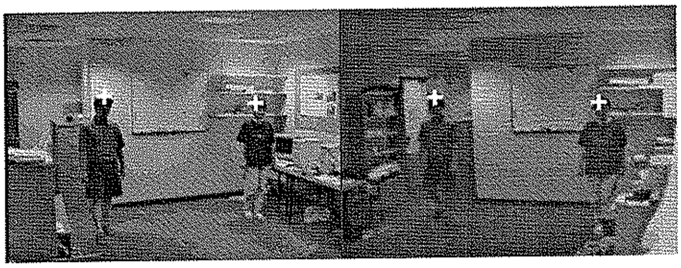


Figure 3.7: *Two images from a sequence pair captured by two widely separately cameras with the extracted highest points of the subjects in the scene.*

3.2.2 Feature Correspondence

Having discussed the preprocessing, the next task is to match the highest points extracted from subject images from one frame to the next. The matching between consecutive frames is generally achieved by searching for the closest match in the subsequent frame based on certain visual features, such as motion (position or velocity) and appearance (size or colour) of the subject. We use the second-order discrete Kalman filter to estimate the *motion vectors*, $\mathbf{z}_x(k) = [x, \dot{x}, \ddot{x}]^T$ and $\mathbf{z}_y(k) = [y, \dot{y}, \ddot{y}]^{\mathrm{T}}$, of the subject's highest point.

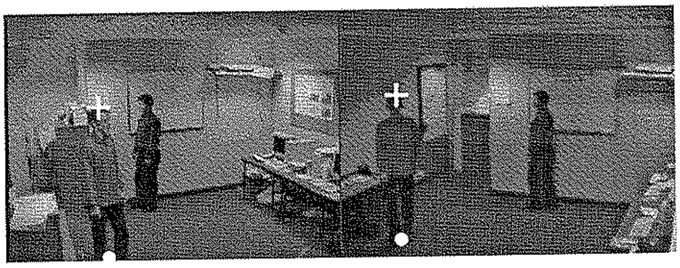


Figure 3.8: *The lowest points of the left subjects in two camera images do not correspond to the same point of the person.*

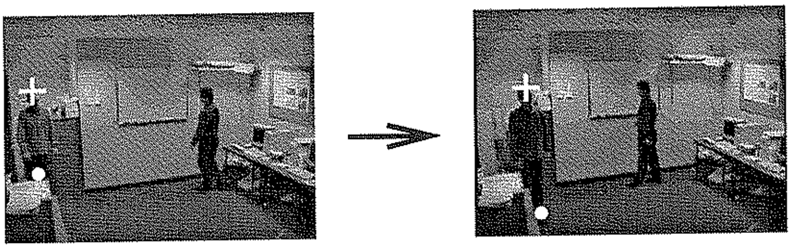


Figure 3.9: *The lowest points of the left subjects in two image frames from a camera do not correspond to the same point of the person.*

The correspondence of the subjects is established based on the predicted motion vectors. This tracking method is related to feature point tracking. A good survey of feature point tracking approaches can be found in [127] and a comparative study of several different schemes with different *linking strategies* (i.e. point correspondence between adjacent image frames) and occlusion handling techniques can be found in [178].

Dynamical Models for Tracking

A Kalman filter is a recursive, linear, optimal data processing algorithm used to estimate the states of a dynamic system in a noisy environment [102]. The estimation algorithm consists of a state prediction stage followed by a correction stage using the measurement. Its recursive nature removes the need to explicitly store a history of all past measurements. This is of vital importance to a practical implementation. The Kalman filter uses a linear system model (Equations (3.2) and (3.3) explained below) to represent the states of a dynamic system with Gaussian state space. Linear systems are desirable in that they are more easily manipulated and stable. The linear system theory is also much more complete and practical than its nonlinear counterpart [102]. However, if the linear models cannot provide satisfactory results, the extended nonlinear filter, extended Kalman filter [9], can be used.

For a linear system with Gaussian distributed stochastic inputs, or noise sources, the Kalman filter is optimal in two senses. First, it is an *unbiased* estimator, so the mean value of the estimated state is equal to the true state. Second, it is an *efficient* estimator (i.e. no other unbiased estimator can have a smaller variance) which produces estimates that minimise the mean-square estimation error between the estimated state and the true state. From a Bayesian viewpoint, a Kalman filter propagates the conditional probability density of the state conditional on the measurements such that the mean, mode and the median of the estimated density all coincide. The Kalman filter has applications throughout computer vision as a general method for tracking, estimation and data fusion given the noise measurement. For example, Rosales and Sclaroff [135] used extended Kalman filters to track multiple moving people. Zhang [191] applied a Kalman filter to the parameter estimation problem for conic fitting. Brown et al. [17] built a system in which object tracking is performed in each sensor node using a Kalman filter to integrate information from all other sensor nodes.

Recently, Isard and Blake [71] have introduced the CONDENSATION algorithm to probabilistically track curves in visual scenes. This algorithm, unlike Kalman filtering which assumes a Gaussian distribution for its stochastic components, uses a set of random samples to represent the propagation of an arbitrary probability density over time. The strength of the CONDENSATION algorithm is its robustness to noise and distractors (e.g. cluttered background) in the image. They have demonstrated that the CONDENSATION algorithm succeeds in tracking the position of a person in agile motion (e.g. dancing) whereas a Kalman filter fails. This is because the algorithm can maintain multi-modal probability distributions, represented by multiple samples, such that multiple hypotheses of the position can propagate over time and cover the agile position. In this case, a uni-modal tracker (e.g. Kalman filter) may fail to track since it can only maintain a single hypothesis of an object position. A multi-modal tracker allows both distractors and the true object position to be represented simultaneously. Once the distractor is found not to satisfy the target dynamics, the true object position will re-assert itself as the probability distribution propagate through the object's dynamic model. However, the CONDENSATION algorithm is computationally expensive in comparison to the Kalman filter. As a consequence, the Kalman filter is used in this thesis for tracking multiple people.

The Kalman Filtering

As noted above, a Kalman filter estimates the true value of the system state vector $\mathbf{s}(k)$ by combining the predicted state $\hat{\mathbf{s}}(k)$ and the measurement vector $\mathbf{z}(k)$. In our case, both the state $\mathbf{s}(k)$ and the measurement $\mathbf{z}(k)$ are the same, i.e. the motion vectors $[x, \dot{x}, \ddot{x}]^{\mathrm{T}}$ and $[y, \dot{y}, \ddot{y}]^{\mathrm{T}}$. Two Kalman filters are used to update the x and y coordinates of the motion vector for each subject assuming that the motion vectors in x and y coordinates are independent of each other. Figure 3.10 shows a block diagram of a Kalman filter. Note that in general the state, $\mathbf{s}(k)$, is not directly observable, so it must be determined through an estimation process to obtain an estimated state $\hat{\mathbf{s}}(k)$. The linear dynamic system (a discrete form) for which the Kalman filter addresses the estimation problem can be described by:

$$\mathbf{s}(k+1) = \mathbf{F}(k)\mathbf{s}(k) + \mathbf{G}(k)\mathbf{w}(k), \tag{3.2}$$

$$\mathbf{z}(k) = \mathbf{H}(k)\mathbf{s}(k) + \mathbf{v}(k). \tag{3.3}$$

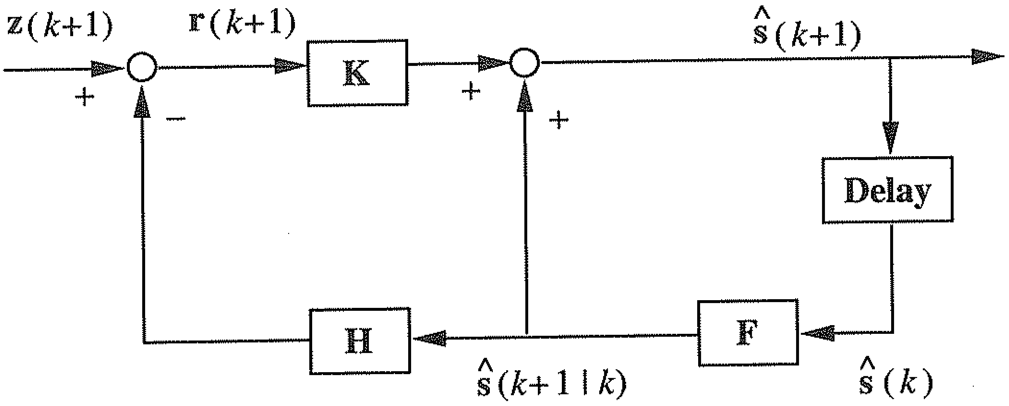


Figure 3.10: *Block diagram of a Kalman filter.*

In the state transition equation (Equation (3.2)), $\mathbf{F}(k)$ is the system model which propagates state over time, $\mathbf{G}(k)$ is the system noise model which accounts for the system noise, $\mathbf{w}(k)$ is the system noise. In the measurement equation (Equation (3.3)), $\mathbf{H}$ is the measurement model which transfers the system state to the measurement space and $\mathbf{v}(k)$ is the measurement noise. The random noise of the system and the measurement are assumed to be independent of each other and modelled as white noise (i.e. uncorrelated over time) with Gaussian distributions:

$$E[\mathbf{w}(k)\mathbf{v}^{\mathrm{T}}(j)] = 0, \text{ for every } k \text{ and } j, \tag{3.4}$$

$$E[\mathbf{w}(k)] = 0, \text{ and } E[\mathbf{w}(k)\mathbf{w}^{\mathrm{T}}(j)] = \mathbf{Q}(k), \tag{3.5}$$

$$E[\mathbf{v}(k)] = 0, \text{ and } E[\mathbf{v}(k)\mathbf{v}^{\mathrm{T}}(j)] = \mathbf{R}(k). \tag{3.6}$$

where $E[\ ]$ is the statistical expectation operator.

Equation (3.4) defines the zero cross-correlation between the two noises. Thus, these two noise values are determined independently. The system noise covariance and measurement noise covariance, $\mathbf{Q}(k)$ and $\mathbf{R}(k)$, are usually assumed to be known and determined on the basis of experience [191]. The estimation steps are as follows:

- Predict the states:

$$\hat{\mathbf{s}}(k+1|k) = \mathbf{F}(k)\hat{\mathbf{s}}(k). \tag{3.7}$$

- Predict the state covariance:

$$\mathbf{P}(k+1|k) = \mathbf{F}(k)\mathbf{P}(k)\mathbf{F}^{\mathrm{T}}(k) + \mathbf{G}(k)\mathbf{Q}(k)\mathbf{G}^{\mathrm{T}}(k). \tag{3.8}$$

- Compute the Kalman Gain:

$$\mathbf{K}(k+1) = \mathbf{P}(k+1|k)\mathbf{H}^{\mathrm{T}}(k+1)[\mathbf{H}(k+1)\mathbf{P}(k+1|k)\mathbf{H}^{\mathrm{T}}(k+1) + \mathbf{R}(k+1)]^{-1}. \tag{3.9}$$

- Compute the *innovation* (i.e. measurement residual):

$$\mathbf{r}(k+1) = \mathbf{z}(k+1) - \mathbf{H}(k+1)\hat{\mathbf{s}}(k+1|k). \tag{3.10}$$

- Update the state estimation:

$$\hat{\mathbf{s}}(k+1) = \hat{\mathbf{s}}(k+1|k) + \mathbf{K}(k+1)\mathbf{r}(k+1). \tag{3.11}$$

- Update the state covariance:

$$\mathbf{P}(k+1) = [\mathbf{I} - \mathbf{K}(k+1)\mathbf{H}(k+1)]\mathbf{P}(k+1|k). \tag{3.12}$$

To estimate the state for time step $k+1$, the filter first predicts the state $\hat{\mathbf{s}}(k+1|k)$ from the previous state $\hat{\mathbf{s}}(k)$ based on the system model $\mathbf{F}(k)$. The previous state covariance matrix $\mathbf{P}(\mathrm{k})$ is used to predict the state covariance matrix $\mathbf{P}(k+1|k)$. This predicted state covariance $\mathbf{P}(k+1|k)$ is then used to compute the Kalman gain matrix $\mathbf{K}(k+1)$ and to update the state covariance matrix $\mathbf{P}(k+1)$. From the measurement matrix $\mathbf{H}(k+1)$, the predicted state $\hat{\mathbf{s}}(k+1|k)$ is transformed to the predicted measurement, $\hat{\mathbf{z}}(k+1|k) = \mathbf{H}(k+1)\hat{\mathbf{s}}(k+1|k)$, for time step $k+1$. The real measurement $\mathbf{z}(k+1)$ is used to compute the innovation $\mathbf{r}(k+1)$. Finally, the innovation is weighted by the Kalman gain $\mathbf{K}(k+1)$ to compute the correction term which is then added to the predicted state $\hat{\mathbf{s}}(k+1|k)$ in order to obtain the estimated state $\hat{\mathbf{s}}(k+1)$.

As has been seen, the purpose of Kalman filtering is to recursively estimate the value of the state by predicting the state based on a system model and using the measurement to correct the predicted state. However, the filter needs an initial estimate to start the estimation process. This is a practical issue of using Kalman filter [16], since the system needs to specify the initial state estimates $\hat{\mathbf{s}}(0|0)$ and state covariance $\mathbf{P}(0|0)$ to start with these specifications in the estimation process. A good initial state estimate ensures fast convergence whereas poor estimates may give rise to slow convergence or even divergence in which case the filter must be re-initialised. Generally, the initial state can be set as $\hat{\mathbf{s}}(0|0) = E[\mathbf{s}(0)]$ or is directly calculated from measurement [134]. The implementation of the Kalman filter including the initialisation is described in Appendix B.

In Kalman filtering, a divergence problem may occur when the system in not observable. This problem is often referred to as the *observability problem* [19], Physically, this means that from the measurements there are one or more state variables that are hidden (unobserved). As a result, the corresponding estimation errors will be unstable. This problem is due to the fact that sometimes the measurements do not provide enough information to estimate all the state variables of the system. Some of the validation tests of the Kalman filter regarding this observability problem can be found in [33]. In this thesis, since the system state is designed as the same as the measurement in the Kalman filter, the measurement matrix is a square regular matrix (see Equation (B.6)). In this case, the solution to the observability problem is trivial as the state and the measurement are the same [13].

Data Association with a Nearest Neighbour Approach

Another issue in applying Kalman filters for tracking multiple objects is the problem of *data association* which is often known as the *motion correspondence problem* [35]. At first sight, tracking might seem to be a special case of an estimation problem. However, it is wider in scope, namely, in addition to the need to use the estimation tools it also requires the use of statistical decision methods when considering the issue of data association [9]. The process of data association is essential to link the measurements to the estimation mechanism when tracking multiple objects [130].

In our case, Kalman filters are used to predict the motion vector of each individual when tracking multiple people. These predictions are then matched to the actual measurements (i.e. motion vectors of the subjects' highest points) in the subsequent frame. At this matching stage, ambiguity may arise. This ambiguity can be seen from the state update (Equation (3.11)) and the innovation computation (Equation (3.10)) where a single measurement $\mathbf{z}(k+1)$ is required to match each of the predicted states $\hat{\mathbf{s}}(k+1|k)$ of different filters for updating the state estimate $\hat{\mathbf{s}}(k+1)$. Since there are multiple subjects to track, the system needs to perform data association for selecting a candidate matching subject in the next frame to represent the subject being tracked.

To perform data association, one possible method is to use qualitative motion heuristics to constrain the candidate matching object [177]. Such methods usually convert qualitative descriptions (e.g. smoothness of motion and rigidity) into quantitative measures and define a distance term for the optimal motion. A threshold is used to identify a valid candidate match (i.e. an object in one image is assigned an object in the other image), whilst a zero distance makes a correspondence optimal. As mentioned at the start of Section 3.2.2, the tracking in SCT is related to the feature point tracking; a good survey of feature point association approaches based on the qualitative motion heuristics can be found in [127].

The other alternative is to build the correspondence based on probability criteria. This thesis uses the *Mahalanobis Distance* (MD) of the predicted motion vectors to estimate the likelihood of data originating from a specific subject. The simplest approach to this problem of associating uncertainty is the nearest neighbour method. Although this method is simple and computationally inexpensive, it has some drawbacks. Alternative approaches can be found in [35, 130]. From a set of candidate matching subjects, the nearest neighbour method selects a single subject that has the closest feature vector. The closest is usually defined using the MD. Furthermore, the nearest neighbour method makes assignment decisions based solely on the current image frame. Better matching results can be obtained by using information from multiple frames, e.g. a track-splitting filter postpones the decision process for using information from multiple frames [35]. Alternative approaches include all-neighbours methods, such as the Joint Probabilistic

Data Association Filter (JPDAF) and Multiple Hypotheses Tracking (MHT) methods. In these all-neighbours methods, multiple measurements can be associated to a single filter, and the match of all candidates are considered jointly [128]. These alternative approaches can perform data association more reliably but have some drawbacks. For example, a track-splitting filter and MHT have exponential complexity. On the other hand, the JPDAF is only applicable to tracking scenarios where the number of targets to be tracked is known [35]. Since this work focuses on matching subjects across cameras, the nearest neighbour method is used in SCT for matching subjects between consecutive frames. Related association problems in the matching for MCCT will be considered in Chapter 4.

Now let us explain how a filter performs a nearest neighbour match to associate subjects between two consecutive frame images. This method assumes that only one measurement can be attached to a filter and a measurement cannot be matched with more than one filter. This method is a likelihood method where the likelihood of a point being the correct match for a filter is defined as the normalised innovation [130]. The normalised innovation, $\mathcal{M}_m$, is the MD between a measurement $\mathbf{z}(k+1)$ and the predicted measurement $\hat{\mathbf{z}}(k+1|k)$ [128, 9]:

$$\mathcal{M}_m = \mathbf{r}^{\mathrm{T}}(k+1)\mathbf{S}^{-1}(k+1)\mathbf{r}(k+1), \tag{3.13}$$

where the subscript, $_m$, in $\mathcal{M}_m$ is indicative of that it relates to a motion-based modality,

$$\mathbf{r}(k+1) = \mathbf{z}(k+1) - \hat{\mathbf{z}}(k+1|k) \tag{3.14}$$

$\hat{\mathbf{z}}(k+1|k) = \mathbf{H}(k+1)\hat{\mathbf{s}}(k+1|k)$, and $\mathbf{S}(k+1)$ is the covariance of the innovation representing the uncertainty between the true measurement and the predicted measurement:

$$\mathbf{S}(k+1) = H(k+1)\mathbf{P}(k+1|k)H^{\mathrm{T}}(k+1) + \mathbf{R}(k+1). \tag{3.15}$$

In the feature space, the points of a given MD form the surface of a d-D ellipsoid where d is the dimension of the measurement vector $\mathbf{z}(k)$ [35]. The diagram on the left in Figure 3.11 shows an example of the points (as a ellipse) of a given MD in 2D feature space where the shading represents the probability density of the measurements about their predicted value. The darker the region, the more likely the correct measurement is to be found. The probability distribution of measurements is highest about the predicted mean value of the measurement and monotonically decreases with increasing distance from the predicted mean value. The diagram on the right shows the measurement of

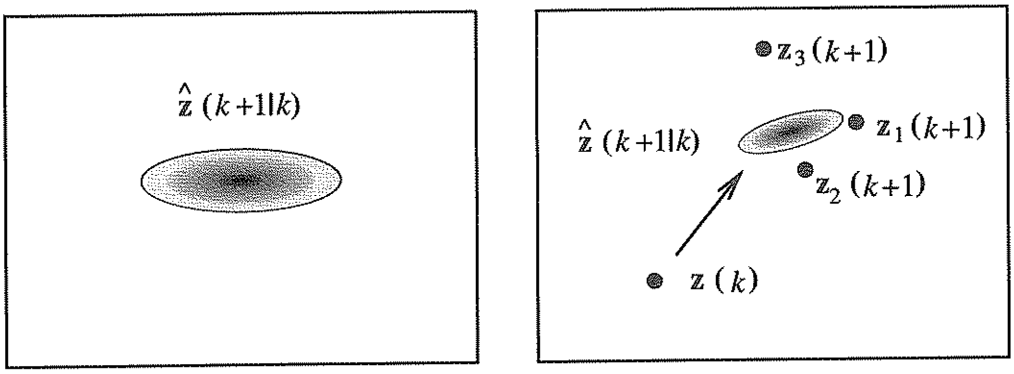


Figure 3.11: *Data association using the nearest neighbour method based on Mahalanobis Distance (MD).*

a target $\mathbf{z}(k)$ at time k and three measurements at time $k+1$. The system associates $\mathbf{z}_1(k+1)$, with the shortest MD with respect to the predicted measurements $\hat{\mathbf{z}}(k+1|k)$, to the $\mathbf{z}(k)$.

To perform the nearest neighbour method, the filter selects the blob with the motion vector which minimises the $\mathcal{M}_m$. Thus, Equation (3.13) is evaluated for the highest point of each blob in the subsequent frame and the blob whose highest point produces the smallest normalised innovation is selected as the best match. The right diagram in Figure 3.11 illustrates the association between the predicted measurements and the true measurement using the nearest neighbour method. The assumption of the nearest neighbour method is that the correct match is more likely to satisfy the nearest neighbour test (i.e. with the smallest $\mathcal{M}_m$) in the feature space. The system then tracks the subjects in each camera based on Kalman filters with the nearest neighbour association method.

3.3 Tracking Results

Figures 3.12, 3.13 and 3.14 illustrate the measured motion vectors $\mathbf{z}_x(k+1)$ and the predicted motion vectors $\hat{\mathbf{z}}_x(k+1|k)$ of a person's highest point for a sequence where the point is extracted from the left subject in the left camera image of Figure 3.7. These measured motion vectors are computed from the extracted highest point of the tracked subject's image. The image sequence contains 590 frames and the subject is visible beginning from the 32^{nd} frame. The prediction results are only shown for the x coordinate since the results for y coordinate are similar. The bottom graph in each figure shows the prediction error of the filter. Note that the filter can follow the highest point well for the scenario when people walk in the office. The mean square errors of predicted position, velocity and acceleration over the whole sequence are 5.46, 1.37 and 1.65 respectively. The maximum prediction error occurs at around the 333^{rd} frame when the person change direction suddenly. This sudden change can result in the filter following the wrong subject if there are other subjects around. The situation can be more serious if there are no measurements (e.g. during occlusion) to correct the predicted states. To handle this situation, this thesis uses two cameras to obtain unambiguous visual information in one camera for solving the occlusion problem in the other camera (see an example in Section 7.3).

Figure 3.15 shows an example of the tracking results in each camera. The system assigns a label as an identity to each detected subject in each camera and keep tracking the subjects with their identities. The correspondence of the subjects between consecutive frames of a camera is based on the predicted motion vectors using Kalman filters. To track people with two cameras cooperatively, the system needs to determine the correspondence of the subjects across camera images (e.g. subject 1 in the left camera image corresponds to subject A or subject B in the right camera image). This inter-camera subject correspondence problem will be addressed in the following chapters.

For the whole sequence, the mean values of the motion vector $[x, \dot{x}, \ddot{x}]^{\mathrm{T}}$ and $[y, \dot{y}, \ddot{y}]^{\mathrm{T}}$ in x and y coordinates are (216.3, 2.80, 1.67) and (200.20, 2.33, 1.75) with standard deviations of (75.44, 3.20, 2.58) and (20.27, 2.38, 3.78) respectively. The covariance matrix for the combined motion vector, $[x, \dot{x}, \ddot{x}, y, \dot{y}, \ddot{y}]^{\mathrm{T}}$, is obtained as:

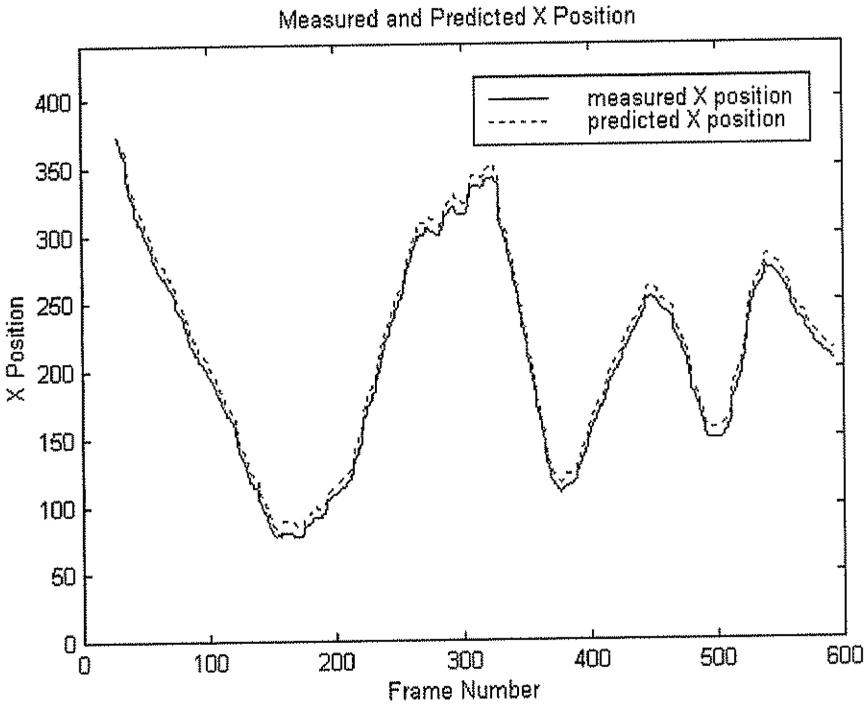
$$\begin{pmatrix} 5691.8 & -17.3 & -4.90 & 4.01 & -1.40 & -1.80 \\ - & 10.23 & 1.39 & 1.98 & 0.40 & -0.36 \\ - & - & 6.65 & -0.03 & 0.60 & 0.96 \\ - & - & - & 410.80 & 11.99 & -4.13 \\ - & - & - & - & 5.69 & 3.13 \\ - & - & - & - & - & 14.26 \end{pmatrix}. \tag{3.16}$$

From this matrix, it can be seen that the covariances between the motion vectors in x and y coordinates are small. For example, the covariance between $\ddot{x}$ and $[y, \dot{y}, \ddot{y}]^{\mathrm{T}}$ is (-0.03, 0.60, 0.96) which indicates that $\ddot{x}$ is largely independent of the motion components in the y coordinate. The independent relations between $\dot{x}$ and $[y, \dot{y}, \ddot{y}]^{\mathrm{T}}$ can also be observed from the matrix where the covariances are small (i.e. (1.98, 0.40, -0.36)).

Comparison of covariance in the matrix also indicates that x is less independent from $[y, \dot{y}, \ddot{y}]^{\mathrm{T}}$ than $\dot{x}$ and $\ddot{x}$ where the covariances are (4.01, -1.40, -1.80). For convenience, this thesis assumes that the motion vectors in x and y coordinates are independent of each other. Two Kalman filters are used to update the x and y coordinates of the motion vector for tracking the highest point of each person.

3.4 Summary

This chapter described tracking using a single camera. Since this research focuses on matching subjects across cameras for tracking people using two cameras cooperatively, only the motion modality is used for tracking in each camera. The method for detecting tracking ambiguity (i.e. questionable matches) and handling the complexity problem in correspondence (i.e. to reduce the number of candidate matches before matching is performed) in SCT (Single Camera Tracking) will be discussed in Section 4.4.3. The following Chapters (4, 5 and 6) present the methodology for matching subjects across cameras in order to perform MCCT (Multiple Camera Cooperative Tracking). The experimental results are given in Chapter 7. The next chapter describes a statistical framework that is used to adaptively fuse all modalities in order to improve the inter-camera subject correspondences.



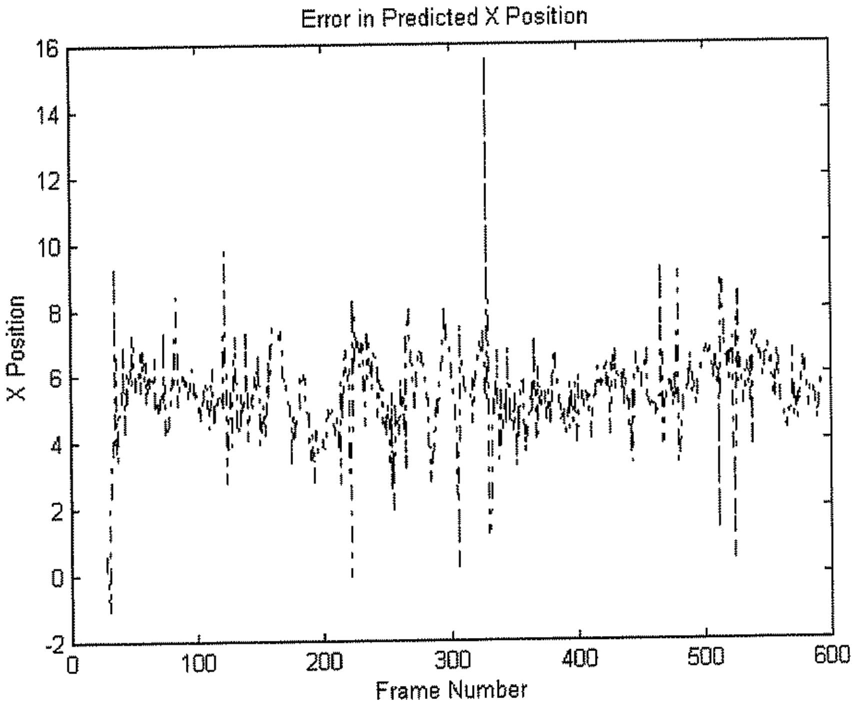


Figure 3.12: *Measured X position and predicted (using a Kalman filter) X position of a person's highest point in a sequence.*

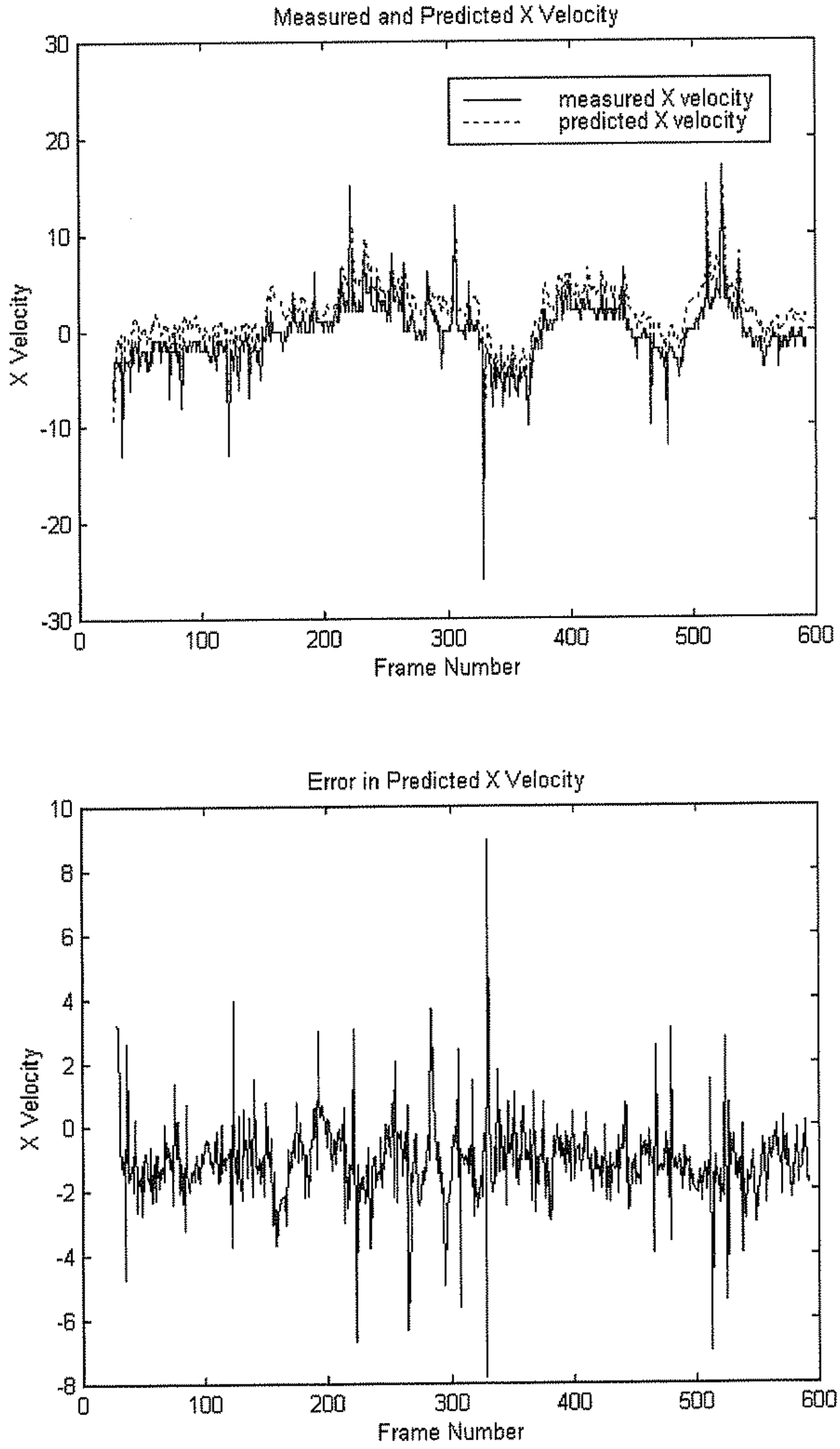
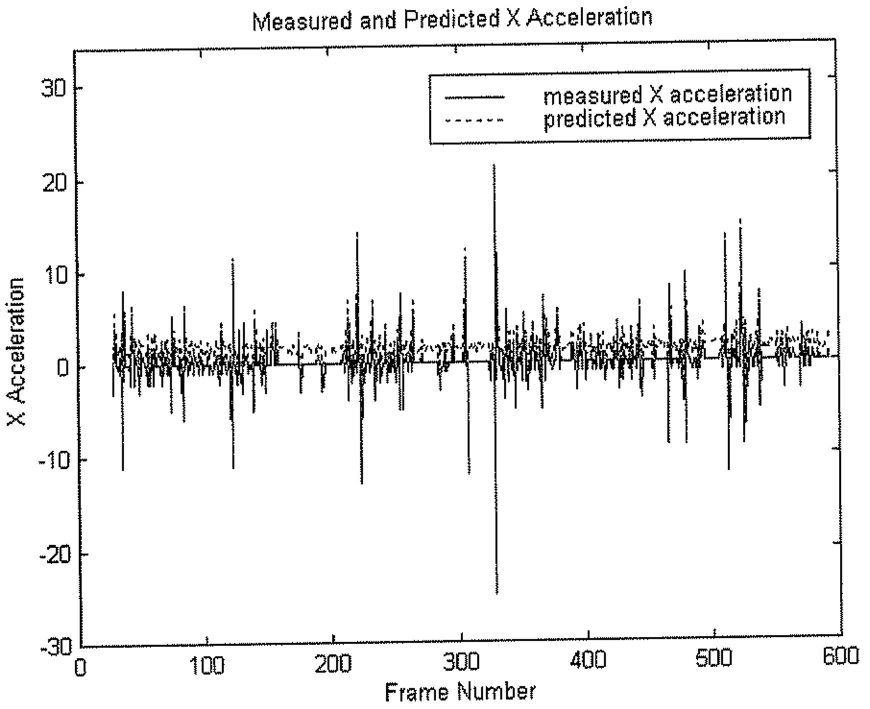


Figure 3.13: *Measured X velocity and predicted (using a Kalman filter) X velocity of a person's highest point in a sequence.*



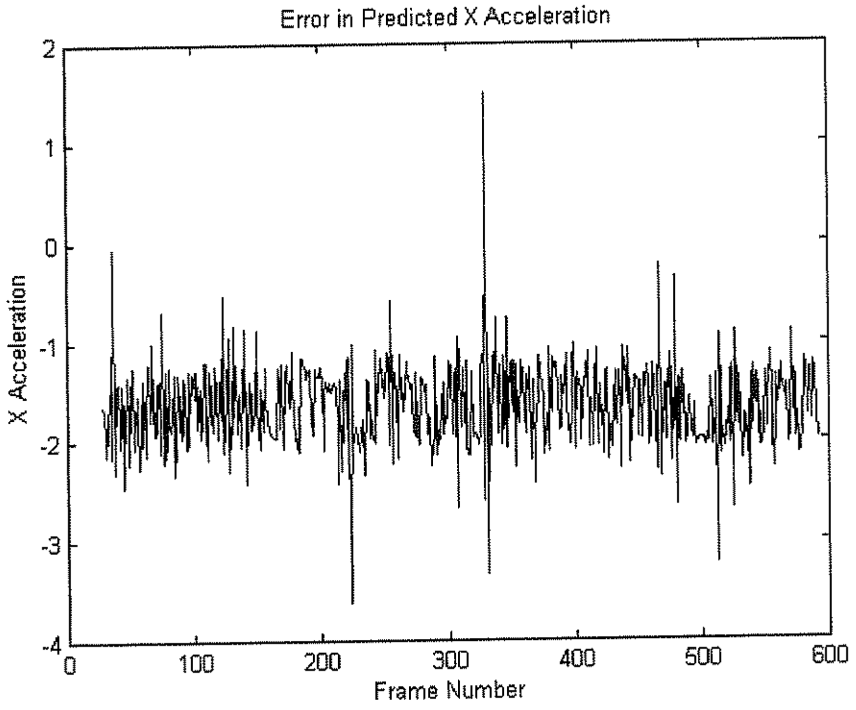


Figure 3.14: *Measured X acceleration and predicted (using a Kalman filter) X acceleration of a person's highest point in a sequence.*

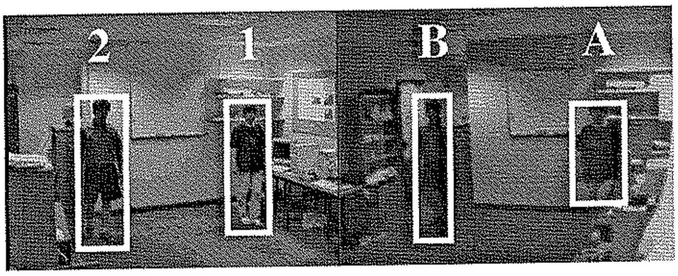


Figure 3.15: *The goal of MCCT is to determine the correspondence of the subjects' images between two camera images.*

Chapter 4

Bayesian Modality Fusion for Correspondence

4.1 Introduction

The previous chapter has described how the system tracks multiple people using a single camera. This chapter focuses on the development of a framework to adaptively fuse multiple visual modalities for matching subjects across cameras in order to perform MCCT (Multiple Camera Cooperative Tracking). The aim of MCCT mode is to pass subject identities (i.e. assigned labels) between cameras in order to track multiple people using two cameras cooperatively. Figure 4.1 shows the block diagram for matching subjects across cameras. This involves two steps: preprocessing and matching the subject images in two camera images. The two-camera system used in this thesis is different from the traditional stereo vision techniques which usually with two cameras closely placed [6]. Since the cameras used in this thesis are widely separated, the images from two cameras have large variation.

Two stages of preprocessing are performed before the matching starts: (1) segmenting moving subjects from a still background and (2) extracting features from the segmented subjects' images in both cameras. The first stage is described in Section 3.2.1 (see Figures 3.5 and 3.6) and the second stage will be explained in Chapter 5 and Chapter 6 where different features modalities (e.g. apparent colour) are used for inter-camera correspondences. After preprocessing, the system begins the matching process by establishing the feature correspondence between two camera images. Instead of tracking with a single camera, the system can pass the subject identities between cameras to keep

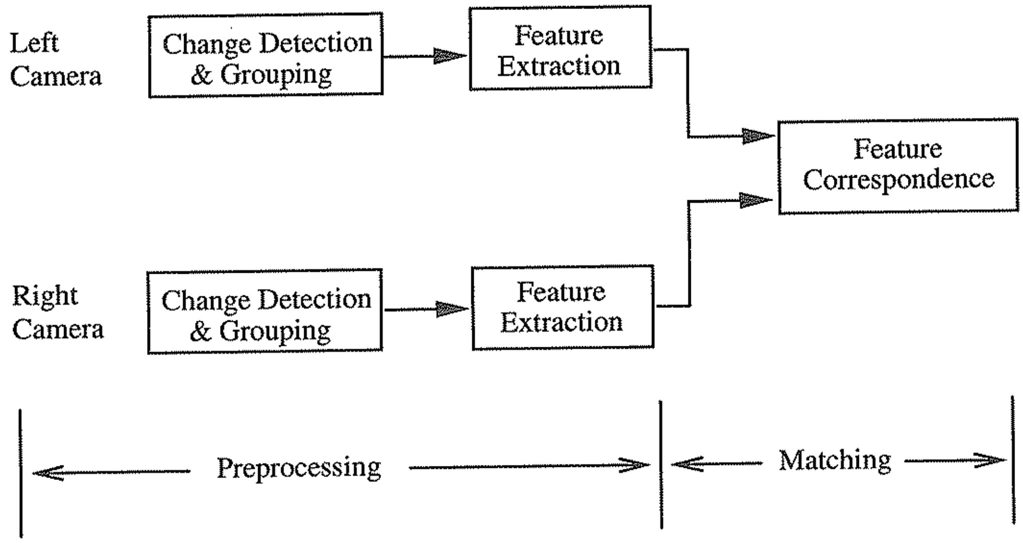


Figure 4.1: *Block diagram of matching the subjects' images between two camera images.*

tracking people even when occlusion occurs in a camera. To pass the subject information across cameras is achieved by matching subjects across cameras. Thus, tracking people using two cameras cooperatively can be more reliable. This chapter will focus on the development of a framework for the matching task in order to effectively fuse multiple feature modalities. These different modalities used for matching subjects across camera images will be described in the following two Chapters 5 and 6.

The major difference between the feature correspondence in SCT (Single Camera Tracking) and MCCT is that in MCCT the image features to be matched are obtained from different cameras' images whereas in SCT the image features are obtained from a single camera. The matching process in MCCT is related to the stereo correspondence problem while SCT is related to the motion correspondence problem. From the geometric point of view, if two identical cameras are used in MCCT, the matching problem in MCCT is the same as the problem in matching images from a moving camera. However, as discussed in Chapter 1, compared to the SCT, the features to be matched in MCCT are extracted from different camera coordinate systems and from different physical processes which make the matching more difficult. Furthermore, the two cameras used are not calibrated so the 3D world coordinates of the subjects are unknown thus making the matching more difficult.

The novelty and main purpose of this chapter is to apply a discrete BBN (Bayesian Belief Network) to fuse multiple modalities based on different features for solving the inter-camera correspondence problem [27, 26]. Previous researchers either have not explained how to fuse [140, 115] or have assumed different modalities are independent [22, 23] (as discussed in Section 2.3). This chapter first discusses the inter-camera correspondence problem in Section 4.2, then introduces the theory of probabilistic inference in the BBN (i.e. computing the probability distributions over a particular subset of random variables given the states of some other variables in the network) in Section 4.3. Finally, the explanation of the use of a BBN to fuse multiple modalities for solving the inter-camera correspondence problem is given in Section 4.4.

4.2 Inter-Camera Correspondence Problem

This section reviews the constraints used for the correspondence problem and explains some of the general issues as well as the approaches taken towards solution in this thesis. The problem of establishing subject correspondences between two camera images is then defined. The task of establishing correspondence is performed by searching for the features in different images which correspond to the same entity in the world. In the literature, a large number of algorithms have been implemented with different types of features, match constraints and search algorithms. Some good reviews can be found in [100, 80, 18].

4.2.1 Feature-Based Constraints for Correspondence Problem

To establish feature correspondence between two images, different constraints based on different features can be used to limit the number of candidate matching features. The constraints for correspondence can be generally divided into two forms: local and global constraints [100]. Local constraints are specific to each individual match (i.e. assigning one feature in an image to another feature in the other image) whereas global constraints are related to the global consistency of multiple or all matches. The normal strategy is to apply the local constraints in the first stage for each feature in an image to identify a set of candidate matching features in the other image. The global consistency (or compatibility) of the local matches is then used to test these local matches to see if each pair of matches is mutually compatible.

Figure 4.2 illustrates an example where global constraint is necessary for evaluating all local matches in order to obtain global consistency. Considering on the 2D image plane, the total number of subjects in both camera images is assumed to be known as three. The features extracted from subjects' images are used to determine the subject

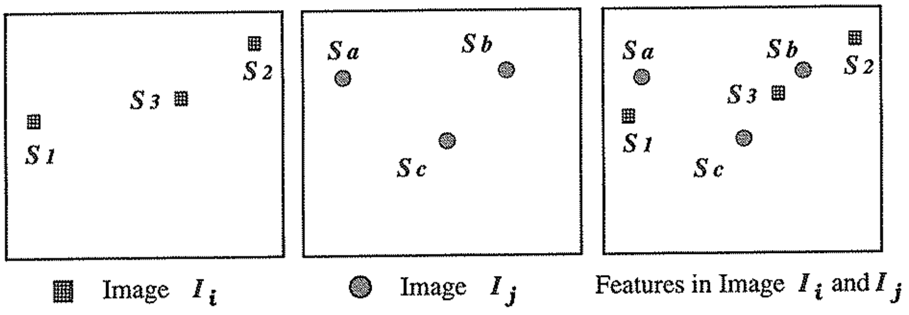


Figure 4.2: *A global constraint is necessary to avoid the conflicting matches.*

correspondences between two images. A match between subjects (i.e. to assign a subject 1, S_1, in one image to another subject, S_a, in the other image) is denoted by $S_1 \leftarrow S_a$. Subjects S_1, S_2, and S_3 in the first camera image, I_i, need to be matched to subjects S_a, S_b, and S_c in the second image, I_j. Among S_a, S_b, and S_c, S_1 corresponds to S_a as a match $S_1 \leftarrow S_a$, where the match is determined based on the Euclidean metric distance measures in the feature space. If matching independently, the best match for both S_2 and S_3 is S_b. Obviously, this pair of matches (i.e. $S_2 \leftarrow S_b$ and $S_3 \leftarrow S_b$) conflict with each other. Evaluating globally by considering this pair of matches jointly, the best match for S_3 is S_c since the combination of assignments $\{S_3 \leftarrow S_c$ and $S_2 \leftarrow S_b\}$ is better than $\{S_3 \leftarrow S_b$ and $S_2 \leftarrow S_c\}$ based on the Euclidean metric distance. This problem of consistency in matching subjects in two images was also considered by the authors of the KidsRoom tracking system [70], as discussed in Section 2.2.1. In the following, a brief review of these two forms of constraints (see Table 4.1) and the explanation of how to apply these constraints to the subject correspondence problem in MCCT are given.

Table 4.1: *Constraints for feature correspondence (see text for explanation).*

Local Constraints	Global Constraints
similarity	uniqueness
epipolar	continuity
disparity gradient limit	topological

Local Constraints

A straightforward local constraint for identifying correct matches is to test the similarity of the attributes of the features, often referred to as a *similarity constraint* [100].

The similarity constraint calculates the disparity (difference between the attributes of the features) and compares the calculated value with a pre-defined threshold. If the value of disparity is smaller than the threshold, they are treated as similar or compatible features.

The other local constraint between two stereo images of the same scene is based on the *epipolar geometry*. This geometry is the only geometry between two stereo images without further assumption, e.g. the homography assuming all points are coplanar (more details in Section 5.2). It is obtained as a result of the imaging geometry of a two-camera system and can be used to limit the search space from 2D (i.e. image plane) to 1D (i.e. the *epipolar line*). This geometry is widely used in traditional stereo vision for finding the corresponding image features. More discussion of the epipolar constraint is given in Chapter 5.

Another local constraint is the *disparity gradient limit* constraint. In this constraint, the disparity of matched points is usually defined as the difference in their pixel positions. The idea is to use the disparity gradient rather than the disparity magnitude to constrain the candidate matching features. For any pair of matches (each match consists of one feature in each of the two images), the disparity gradient is defined as the ratio of the difference in disparity of the two matches to the average disparity of the matches between two images. Thus the limit on the disparity gradient between neighbouring matches can be used to constrain the correspondence. This type of constraint is also used in stereo vision, e.g. matching points along line segments [179]. However, this type of constraint may be computationally expensive due to the need to calculate average disparity of the matches between the two images. The aim of inter-camera subject correspondences is for tracking people using two cameras cooperatively. The tracking should be performed in real-time or at least as quickly as possible. As a consequence, the constraint used for determining subject correspondences needs to be less computationally expensive. For example, in applying this constraint to match a person's highest point in two cameras, the calculation of average disparity would have to take account of a greater number of points. Thus, the disparity gradient limit constraint is not considered in this research.

Global Constraints

Among different global constraints, the *uniqueness constraint* is the most straightforward but also the most general [100]. This constraint requires each item in an image to be assigned to one and only one matching candidate in the other image. It is simple but widely used in many matching strategies, e.g. [166]. This global constraint is adopted for inter-camera subject correspondence, assuming people are in the overlapping FOVs of two cameras. Note that when not all of the people in the office are in the overlapping FOVs of two cameras, this uniqueness constraint may result in incorrect matches. In this case, this is because subjects in two camera images do not necessarily correspond to the same people in the 3D world.

Continuity constraint is another global constraint which depends on the observation that points adjacent in 3D space remain adjacent in each image projection. This constraint can be applied in different ways. For example, neighbouring edge points should have "similar" disparity values, or the connected edge points in one image must match to connected edge points in the other image. However, the inter-camera correspondence problem considered in this thesis requires people to be matched across cameras. The image structures to be matched are different separated blobs in two camera images. The suitable image structures for applying continuity constraint are in general adjacent points in 3D space. As a consequence, this constraint may be more applicable to matching geometric features (e.g. points) of a single object. Since our goal is to match multiple different subjects, this constraint is not suitable for our problem.

Another popular global constraint is the *topological constraint*, such as the *relative position constraint*, which is based on the assumption that 3D structure viewed in both images being identical. However, this constraint may only hold for stereo cameras with a short baseline. In MCCT using two widely separated cameras, the relative positions of people with different depth in different camera images can be dissimilar due to the parallax (induced when parts of the scenes have differences in depth which can be caused by camera translation) [80]. Figure 4.3 shows an example of this case where the relative positions of subjects' images in two camera images are different. Thus, topological constraint is not considered in this thesis.

Approach to Adopting the Constraints

For solving the inter-camera subject correspondences in MCCT, the uniqueness global constraint is used to first limit the number of possible matches based on the global consistency. The details are explained in Section 4.2.3. The local constraints of all visual modalities used for matching subjects across cameras are formulated using a Bayesian framework. Instead of comparing the attribute values of features directly, we adopt two methods to compensate for the feature variations for making the inter-camera subject correspondences more reliable. These two methods are related to two categories

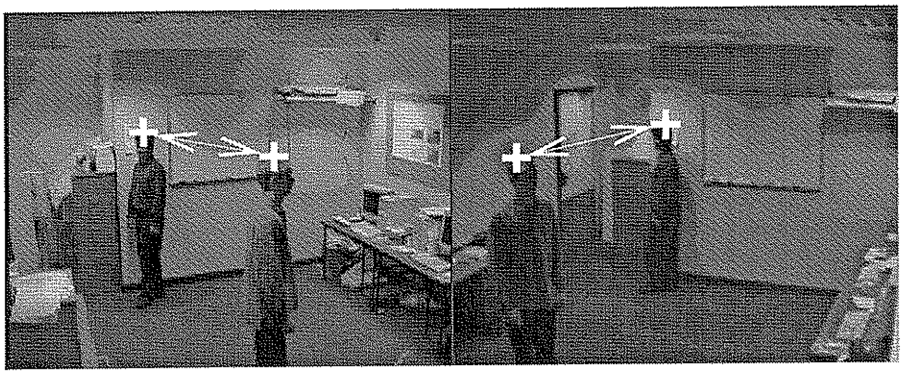


Figure 4.3: *The relative positions of two subjects are very different in two camera images.*

of modalities for matching subjects across cameras, which are:

- geometry-based modalities
  using multi-view geometry (i.e. homography and epipolar geometry) and the scene landmarks to estimate the geometric positional relationship of the corresponding subjects between two camera images (Chapter 5), and
- recognition-based modalities
  using the learnt mapping of the visual information (apparent height and apparent colour) between two camera images to estimate the subject appearances across camera images (Chapter 6).

4.2.2 Searching for the Unique Correspondences

An important issue in matching subjects in two camera images is data uncertainty. Besides the inherent uncertainty in the data provided by sensors, a dynamic scene may be complex and cluttered. Moreover, due to the ambiguous positions of extracted features, the features used for matching can be unreliable (see an example in Figure 5.10). As a consequence, the multiple modalities of different features can be less reliable or even conflict with each other. This phenomenon subsequently results in inconsistent matches. A single constraint is usually not powerful enough to locate all the matches uniquely and correctly. Therefore, the use of multiple modalities of different constraints is necessary to make matching more robust.

Since different modalities can have different reliability, different modalities should be combined with adaptive weights according to image context information in order to efficiently fuse different information. On the other hand, these different modalities regarding different constraints are usually highly correlated since they are all related to the same visual scene. Simply assuming that all modalities are independent ignores such correlation and subsequently causes the scene to be interpreted less reliably.

Searching Approach

In order to reliably infer a unique correspondence for each subject in two images (i.e. a final set of matches between subjects in two images), a framework is required to handle data uncertainty, adaptively fuse multiple modalities and capture the correlation between different modalities and the feature correspondence in two camera images. Bayesian Belief Networks (BBNs) [120, 76] are adopted for these purposes. A BBN is used to model dependencies between modalities and subject correspondences between the two camera images. BBNs also enable the full set of possible matching assignments to be simultaneously considered in a consistent and probabilistic manner in order to infer a unique correspondence for each subject in two camera images. This method of adaptively fusing multiple modalities using BBNs was introduced as *Bayesian modality fusion* [171], where the task was to match a single object (i.e. tracking a human head) between image frames from a single camera. This thesis extends this method for matching multiple subjects between image sequences from two widely separated cameras.

Alternative Searching Approaches

There are alternative approaches for searching for a unique correspondence between two images in the stereo matching literature. In general, these searching algorithms can be classified into two categories as follows [100].

- Relaxation labelling methods which group the information of the neighbouring features iteratively to update the match probability.
  From the iteration computation, this algorithm incorporates the total visual evidence provided by all labelled features. However, the use of a recursive search methods can be computationally very expensive [80].

- Hierarchical schemes which usually perform matching at different levels.
  Coarse features are first matched, and the results are then used to guide the matching of finer features. These methods are generally more appropriate for the structural matching problem. For our problem, the image structure of the positions of multiple subjects in two widely separated cameras are not necessarily the same. Therefore, this type of matching strategy is not considered.

Moreover, the correspondence problem can also be cast as an optimisation problem by minimising a cost function that measures the matching error. The disadvantage to this method is that it generally requires a very high computational cost [80]. In summary, these alternative methods cannot effectively handle data uncertainty and capture the correlation between variables in the application presented in this thesis.

Another important issue in the correspondence problem is computational complexity which cannot be ignored in tracking [127]. In order to cope with this problem, the system needs to reduce the number of candidate matches before matching is performed. For the case where there are m subjects in two images, the uniqueness constraint is used to reduce the complexity from $O(m^m)$ to $O(m!)$ (explained below) before applying a BBN to infer a unique correspondence for each subject in two images. Other methods to address this complexity issue are discussed in Section 4.4.3.

4.2.3 Problem Definition

Let us now define formally the inter-camera subject correspondences problem. Firstly, the maximum number of subjects in each camera image is constrained to be m. Note that the number of subjects in two images are unknown and not necessarily the same. In order to handle the consistency issue in matching (i.e. for avoiding conflicting matches, as discussed in Section 2.2.1), the matching problem is considered as follows. To match m subjects in two images, I_i and I_j, instead of matching each single subject independently with the possibility of conflicting results, all matches are evaluated globally. To globally consider matches for all subjects, a combination of assignments is defined as a union of m matches. Each match assigns a subject in I_j to a subject in I_i. Thus, in a combination of assignments A_α, all m subjects in I_i are assigned to a subject in I_j respectively. After the theory of BBNs is introduced in Section 4.3, the situations when the subject numbers in two camera images are unequal and less than m will be discussed in Section 4.4.1. For m subjects in each of the two images, there are totally m^m possible assignment combinations. After applying the uniqueness constraint (i.e. in A_α, a subject in I_j is allowed to be assigned to one and only one subject in I_i), there could be $m!$ possible assignment combinations, $A_\alpha = \{A_1, \cdots, A_{m!}\}$. Given the visual evidence e of different modalities from two cameras, which might be uncertain and incomplete, our goal is to find the most likely assignment combination which maximises the posterior:

$$\underset{\alpha \in \{1,\cdots,m!\}}{\operatorname{argmax}} \; p(A_\alpha|\mathrm{e}). \tag{4.1}$$

4.3 Bayesian Belief Networks

In order to search for a unique correspondence between two images this thesis applies a BBN (Bayesian Belief Network) to probabilistically infer the feature correspondence between two camera images to obtain the most likely assignment combination. This section is organised as follows. Section 4.3.1 discusses different approaches to data uncertainty. Section 4.3.2 describes BBNs and their limitations. Section 4.3.3 introduces the graphical model of BBNs. Section 4.3.4 provides the secondary structure of BBNs for dealing with multi-connected networks (see Figure 4.6) which stops messages cycling forever in the original BBNs.

4.3.1 Data Uncertainty

Information gathered from different sensors is often uncertain, incomplete, or even conflicting [1]. To match subject across cameras based on different image features, a system is required not only to extract reliable features but also to deal with the uncertainty in the data from two cameras. The simplest way is to average all modalities for building correspondence without considering the uncertainty problem. However, this procedure may not be suitable for integration of data with extreme dispute [1]. There have been various methods proposed for fusing data from different knowledge sources by representing and propagating uncertainty in expert systems. These methods include non-numerical techniques (e.g. *rule-based* methods which use a set of logical rules) and numerical techniques [76]. Numerical methods (e.g. Bayesian approach, Dempster-Shafer theory and fuzzy set theory) have a different perspective on uncertainty and manipulate uncertain information quantitatively [63]. Each method has advantages and limitations.

In the Bayesian approach, uncertainty is viewed probabilistically. Probability can be interpreted as a relative frequency ranging between never occurring to always occurring. The Dempster-Shafer method is based on the theory of belief (or evidence) where uncertainty is viewed as a degree of belief and the belief ranging between total belief and lack of belief with intermediate values corresponding to partial belief. In fuzzy set theory, uncertainty is viewed as a degree of set membership. This degree ranges between *a member* and *not a member*. The advantage of Dempster-Shafer theory is that the evidence supporting one hypothesis does not necessarily decrease the belief in others, as opposed to probability theory [8]. However, unlike Bayesian theory, the theory of belief does not allow a *priori* knowledge to dominate the inference process. Also, the belief strength can be sensitive to the numerical values of input information [1]. Fuzzy set theory is well suited to applications where the evidence is itself fuzzy in nature. However, although fuzzy set approaches seem to have more flexibility than those in probability and belief theories, their performance for fusing contradictory information is generally unsatisfactory [1]. This is due to the general lack of formal definitions in fuzzy set theory [8]. By comparing these methods, Henkind and Harrison [63] concluded that the Bayesian approach is well suited for applications where some prior probabilities are known, and is an attractive approach because of its strong theoretical foundation.

Most classical inferential models do not permit the introduction of prior knowledge into the evaluation process. For the rigours of a scientific method, this is an appropriate response in order to prevent the introduction of extraneous data that might skew experimental results. However, there are cases where prior knowledge provides a useful contribution to inference. For example, the goal of this thesis is to track people in an indoor environment where the background scene is fairly static and cameras are fixed. Prior knowledge of the scene and system architecture can be used to make scene interpretation more reliable. Therefore, the Bayesian approach is more appropriate for the problem. This thesis adopts a BBN to handle data uncertainty in order to make matching subjects across cameras more reliable. Also, compared to non-numerical techniques, BBNs can more efficiently and correctly represent data uncertainty [120]. This is because BBNs can capture dependencies between different visual modalities and the subject correspondences between two camera images.

4.3.2 Bayesian Belief Network

Bayesian belief networks (also known as *Bayesian nets*, *belief networks* or *causal probabilistic networks*) are graphical models that represent the dependencies embedded in probabilistic models [120]. Graphical models are a marriage of probability theory and graph theory [82]. Fundamental to the idea of a graphical model is the notion of modularity where a complex system is built from a combination of simple parts. Probability theory provides the connection that combines the parts, ensuring that the whole system is consistent. It also provides ways to interface models and deal with data uncertainty. The graph theoretic side of graphical models provides an interface by which humans can model highly-correlated sets of variables and a data structure that lends itself naturally to the design of efficient general-purpose algorithms.

BBNs in Computer Vision

BBNs are attractive for computer vision applications for two reasons. First, BBNs offer the ability to deal with the inherent uncertainty in data provided by sensors sampling a dynamic and complex scene. Second, BBNs combine a natural mechanism for expressing domain knowledge, with efficient algorithms for probabilistic inference and learning [132] (e.g. learning the causal relationships between different variables in order to gain a more reliable interpretation of a problem). From the late 1980s, BBNs began to draw the attention of researchers in computer vision, e.g. geometric modelling [30], perceptual integration [139] and scene surveillance [21]. Lately, BBNs have become a popular tool in computer vision and pattern recognition applications, such as object recognition [126], learning dynamic scenes [110], vision-based speaker detection [132] and visual tracking [149]. Researchers have also examined conceptual links between BBNs, Hidden Markov Models (HMMs) and Kalman filters. Both HMMs and Kalman filters can be represented by BBNs with specific prototypical independencies and repetitive structures over time. This temporally repetitive structure is usually referred to as *dynamic Bayesian Belief Networks* [40]. An example is given by the Lumiere project which considers temporal dependencies between a user's goals at different times and the user's behaviour [65].

Bayesian reasoning and inference procedure have been used in a number of research areas for a long time but have only recently gained popularity in multi-sensor fusion [151]. Kortenkamp [89] was one of the first to propose the use of a BBN for multi-sensor fusion. They used BBNs to build a *topological map* for guiding a robot to explore its environment. This thesis uses a BBN to fuse multiple modalities from two cameras and probabilistically infer the subject correspondences between two cameras in order to track people using the cameras cooperatively.

The Limitations of BBNs

Although BBNs have remarkable power and potential in addressing inferential processes, they have some inherent limitations. One potential problem was evident from the Microsoft's Lumiere project [65], which uses a BBN to understand users actions and questions. The BBN is used to infer and provide intelligent assistance based on the model of user behaviour pre-defined by a human expert. However, the system ignored the possibility that the users might wish to violate the probability distribution (i.e. user behaviour model) upon which the system was built. The possibility of a user making a novel request for information in a previously unanticipated way must be accommodated. This problem results from the heavy reliance of the system on prior knowledge such that it cannot handle some previously unforeseen events. However, this may equally be regarded as an advantage of BBNs since prior or domain knowledge is important for some real-world problems, especially when data is scarce, expensive or incomplete [61]. Another problem is that to calculate the probability of any branch of the network, all branches may also need to be calculated. Also, a BBN is only useful when the prior knowledge is reliable. As a consequence, selecting the proper distribution model to describe the system has a notable effect on the quality of the resulting network. In the following, Section 4.3.3 and Section 4.3.4 introduce the theory of BBNs. The readers familiar with the theory of BBNs are advised to proceed to Section 4.4.

4.3.3 Bayesian Graphical Model

Representation of BBNs

A BBN is a Directed Acyclic Graph (DAG) in which each variable is represented by one node. A causal relationship is indicated by each edge represented as a directed link between variables. Mathematically, this type of structure is called a *directed graph.* Figure 4.4 shows an example of a directed graph where "X is the cause of Y" ($X \rightarrow Y$). This causal relationship indicates that the variable Y is conditionally dependent on X. Node X is called a *parent node* of node Y, and Y is called a *child node* of X. Both Y and Z are the *descendants* of X.

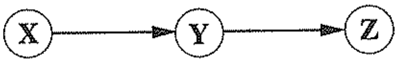


Figure 4.4: *A directed graph.*

Given a set of n variables $\mathbf{V} = \{V_1, \dots, V_n\}$, without knowing the dependencies among variables, one can apply the chain rule of basic probability theory and decompose the joint probability distribution over the variables $\{V_1, \dots, V_n\}$ as:

$$P(\mathbf{V}) = P(V_1)P(V_2|V_1)P(V_3|V_1, V_2) \dots P(V_n|V_1, \dots, V_{n-1}). \tag{4.2}$$

By exploiting the causal relationships between variables, a BBN represents the factorisation of the joint distribution via a sparse set of conditional probabilities [120]:

$$P(\mathbf{V}) = \prod_{i=1}^{n} P(V_i|\Pi_{V_i}), \tag{4.3}$$

where Π_{V_i} is the set of parent nodes of node V_i. If V_i has no parent nodes, $P(V_i|\Pi_{V_i})$ degenerates to the prior $P(V_i)$. Figure 4.5 shows an example of a BBN with the set of variables $\mathbf{V} = \{A, B, C, D, E, F, G, H\}$. The joint probability distribution is decomposed as:

$$P(\mathbf{V}) = P(A)P(B|A)P(C|A)P(D|B)P(E|C)P(F|C)P(G|D,E)P(H|E,F). \qquad (4.4)$$

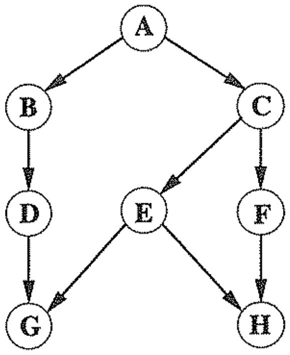


Figure 4.5: *A Bayesian belief network.*

Structure of BBNs

A strength of BBNs is in their representation of probability distributions which can efficiently encode both the independence and dependence relationships among random variables. The independencies can be exploited to provide savings in the representation of a distribution and in computation of the probabilistic inference [12]. In BBNs, a variable is independent from its non-descendants in the network, given the state of its parents and children [120]. Further independent statements that follow from these local statements can be read from the network structure using a graph-theoretic criterion called *d-separation* [120]. Other independence based on context information can be exploited based on the fact that some variables are only relevant in certain contexts [12].

In general, networks can be constructed with continuous or discrete variables. In this thesis, discrete random variables are adopted, where each variable may take on values from a finite set. To construct a discrete BBN, one needs to define both the network topology and the Conditional Probability Tables (CPTs) for each node. The CPT describes the conditional distribution given different assignments of values (or states) on its parent nodes. By specifying the graphical and numerical components of the network, the prior and domain knowledge can be encoded. Some good surveys of learning both structure and parameters (i.e. CPTs) of a network can be found in [61, 119]. Although algorithms exist for automatically structuring a network from training data, BBNs are often constructed by hand. This is the approach adopted in this thesis. For many applications, this should be seen as an advantage rather than a drawback. Since BBNs provide a rich and principled framework for embedding domain knowledge, users may often prefer to specify the network structure and estimate the conditional probabilities associated with the graph edges, $P(V_i|\Pi_{V_i})$.

Probabilistic Inference in Discrete BBNs

By exploiting the encoded independence between variables, an accurate and globally consistent representation of $P(\mathbf{V})$ can be obtained through exact probability propagation in networks. Basically the revision of the global probability distribution (for new observed data in some variables) is decomposed into a sequence of local computations by exploiting the independence properties implied by the model. In the last decade, different exact computations have been proposed to solve probabilistic inference problems formulated by discrete BBNs. Three general approaches are the arc reversal/node reduction technique of Shachter [144], the message passing algorithm introduced by Pearl [120], and the "clique tree" approach of Lauritzen and Spiegelhalter [91]. The Lauritzen and Spiegelhalter algorithm was further developed by Jensen and others to form the basis of the HUGIN expert system shell. Jensen et al. [78] extended earlier work restricted to singly connected trees to cover multi-connected trees (where there may be more than one undirected path between any two nodes, see Figure 4.6) by introducing a *junction tree*, involving a compilation step that transforms a BBN into a secondary structure. This computational approach is adopted in this thesis to handle the multi-connected trees. Previous approaches are not appropriate for multi-connected trees, where messages can cycle forever in the loop. For large BBNs, there are some approximate inference methods (based on the stochastic simulations, e.g. [39]) that provide a better run time.

4.3.4 The Secondary Structure of Bayesian Belief Network

As noted above, this thesis uses a secondary structure of BBNs to handle multi-connected trees in a network. Let us now explain how to obtain $P(\mathbf{V}|\mathbf{e})$ in the context of visual evidence, $\mathbf{e}$, in the secondary structure. Note that the goal of apply BBNs is to obtain posterior $P(\mathbf{V}|\mathbf{e})$ (see Equation (4.1)) given the evidence $\mathbf{e} = \{e_1, \cdots, e_n\}$ in the observed variables. In briefly reviewing the basics of inference algorithms for computing belief

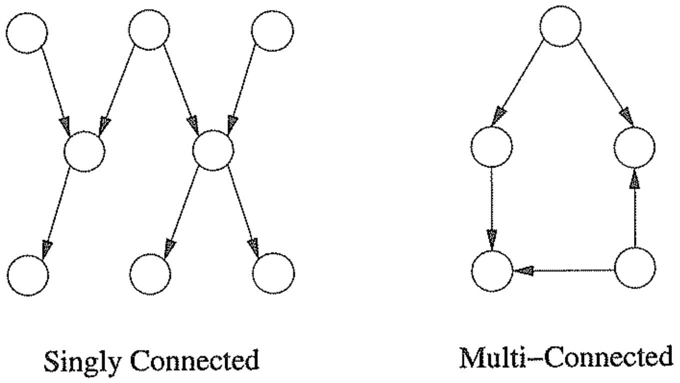


Figure 4.6: *Examples of a singly connected and a multi-connected network structure.*

in the context of observed evidence, a description of the notational conventions and fundamental operations is first given before an introduction to the secondary structures of the BBNs is provided. Finally, inference procedures in the secondary structure for integrating observed evidence are described. We follow Jensen's definitions [76] and Huang and Darwiche's procedural guide [66] to inference in BBNs.

Notation and Algebra

A *variable* denoted with italic uppercase V can have variable values v; and *a set of variables* is denoted with bold uppercase $\mathbf{P} = \{P_1, \cdots, P_n\}$. By assigning a value v to a variable V, v is called an *instantiation* of V. To *instantiate* a set of variables $\mathbf{P}$, a value is assigned to each variable in $\mathbf{P}$ with the assignment $\mathbf{p} = \{p_1, \cdots, p_n\}$ (i.e. a set of values) called an *instantiation* of $\mathbf{P}$.

- *Potential*: A potential $\phi_{\mathbf{P}}$ is a function over $\mathbf{P}$ which maps its instantiation $\mathbf{p}$ into a non-negative real number. $\phi_{\mathbf{P}}(\mathbf{p})$ is called an *element*. A potential can be viewed as a matrix and implemented as a CPT (Conditional Probability Table).

- *Multiplication*: Let $\mathbf{P}$ and $\mathbf{Q}$ be both a set of variables with potentials $\phi_{\mathbf{P}}$ and $\phi_{\mathbf{Q}}$ respectively. The multiplication of $\phi_{\mathbf{P}}$ and $\phi_{\mathbf{Q}}$ is a potential $\phi_{\mathbf{Z}}$ defined as:

$$\phi_{\mathbf{Z}} = \phi_{\mathbf{P}}\phi_{\mathbf{Q}}, \tag{4.5}$$

  where $\mathbf{Z} = \mathbf{P} \cup \mathbf{Q}$ and each $\phi_{\mathbf{Z}}(\mathbf{z})$ is computed as follows:

  1. Identify the instantiations $\mathbf{p}$ and $\mathbf{q}$ that are consistent with $\mathbf{z}$.
  2. Assign to $\phi_{\mathbf{Z}}(\mathbf{z})$ the product $\phi_{\mathbf{P}}(\mathbf{p})\phi_{\mathbf{Q}}(\mathbf{q})$.

For example, let $\mathbf{P} = (A, B)$, $\mathbf{Q} = (B, C)$, $\mathbf{p}_u = (a_i, b_j)$ be an instantiation of $\mathbf{P}$ and $\mathbf{q}_v = (b_j, c_k)$ be an instantiation of $\mathbf{Q}$. If $\phi_{\mathbf{Z}} = \phi_{\mathbf{P}}\phi_{\mathbf{Q}}$, $\mathbf{Z} = (A, B, C)$ and

$$\phi_{\mathbf{Z}}(a_i, b_j, c_k) = \phi_{\mathbf{P}}(\mathbf{p}_u)\phi_{\mathbf{Q}}(\mathbf{q}_v). \tag{4.6}$$

- *Marginalisation*: Let $\mathbf{P}$ and $\mathbf{Q}$ be both a set of variables where $\mathbf{P} \subseteq \mathbf{Q}$, and $\mathbf{Q}$ has the instantiation $\mathbf{q}$. The marginalisation of $\phi_{\mathbf{Q}}$ into $\mathbf{P}$ is a potential $\phi_{\mathbf{P}}$ denoted as follows:

$$\phi_{\mathbf{P}} = \sum_{\mathbf{Q}\setminus\mathbf{P}} \phi_{\mathbf{Q}}, \tag{4.7}$$

where each element $\phi_{\mathbf{P}}(\mathbf{p})$ is computed as follows:

1. Identify the instantiations $\mathbf{q}_1$, $\mathbf{q}_2$, $\cdots$ that are consistent with $\mathbf{p}$.
2. Assign to $\phi_{\mathbf{P}}(\mathbf{p})$ the sum $\phi_{\mathbf{Q}}(\mathbf{q}_1) + \phi_{\mathbf{Q}}(\mathbf{q}_2) + \cdots$.

For example, let $\mathbf{P} = (A, B)$, $\mathbf{Q} = (A, B, C)$ and $\mathbf{p}_k = (a_i, b_j)$ be an instantiation of $\mathbf{P}$, and there are exactly m different instantiations in $\mathbf{Q}$ for which A is instantiated as a_i and B is instantiated as b_j, namely the mutually exclusive instantiations (a_i, b_j, c_1),$\cdots$,(a_i, b_j, c_m). If $\phi_{\mathbf{P}}$ is defined as Equation (4.7),

$$\phi_{\mathbf{P}}(\mathbf{p}_k) = \phi_{\mathbf{Q}}(a_i, b_j, c_1) + \cdots + \phi_{\mathbf{Q}}(a_i, b_j, c_m). \tag{4.8}$$

The Secondary Structure

The secondary structure of a BBN is an undirected tree $\mathcal{T}$ where a node represents a set of variables called a *cluster* $\mathbf{C}$ (instead of a single variable which is the case in the original BBNs). Each edge is labelled with a set of variables which are the intersections of adjacent clusters. This set is called a *separator* $\mathbf{S}$. Figure 4.7 shows the secondary structure obtained from the original BBN in Figure 4.5. It contains clusters $\{ABD, ADE, DEG, ACE, CEF, EFH\}$ and separators $\{AD, DE, AE, CE, EF\}$. The construction of this secondary structure is given in Appendix C.

The joint distribution in Equation (4.3) can also be encoded in the secondary structure of a BBN (see Equation (D.5) in Appendix D) with $P(\mathbf{V})$ defined as:

$$P(\mathbf{V}) = \frac{\prod_i \phi_{\mathbf{C}_i}}{\prod_j \phi_{\mathbf{S}_j}}. \tag{4.9}$$

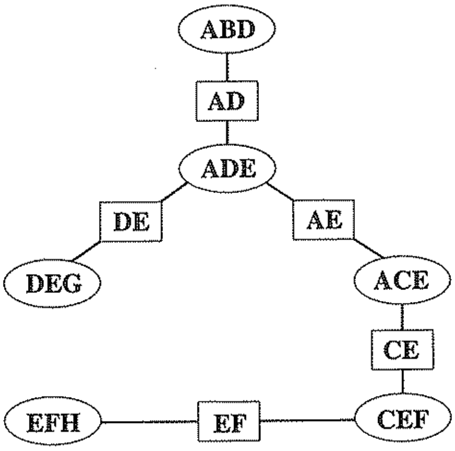


Figure 4.7: *The secondary structure of the BBN in Figure 4.5.*

An important property of the secondary structure is that for each cluster **C** and each separator **S**, it holds that

$$\phi_{\mathbf{C}} = P(\mathbf{C}), \tag{4.10}$$

and

$$\phi_{\mathbf{S}} = P(\mathbf{S}). \tag{4.11}$$

From this property, the probability distribution of any variable V can be computed from any cluster **C** (or separator **S**) that contains V as:

$$P(V) = \sum_{\mathbf{C}\backslash V} \phi_{\mathbf{C}}. \tag{4.12}$$

Probabilistic Inference in the Secondary Structure

Having described the structure of the junction tree, we are now concerned with computing the probability distribution of a variable V given evidence **e**, $p(V|\mathbf{e})$, in a secondary structure. Figure 4.8 illustrates the overall control for the inference procedures. The BBN (including structure and the CPTs) is designed off-line. AT run time, the BBN is then transformed to the secondary structure and initialised to make the BBN to represent the joint distribution of all variables (see Equation (D.5)). After initialisation, the BBN is ready to enter the observations. The dotted path indicates the control of inference procedures with dynamic observations. From this dynamic observations, the previous observations can be considered in the inference process. After marginalisation and normalisation, the network is then re-initialised to incorporate new observations.

The BBN used in this thesis, the posterior of the subject correspondences inferred in the last frame is used as the prior probability in the correspondence node in the current frame. The details of inference are described in Appendix D.

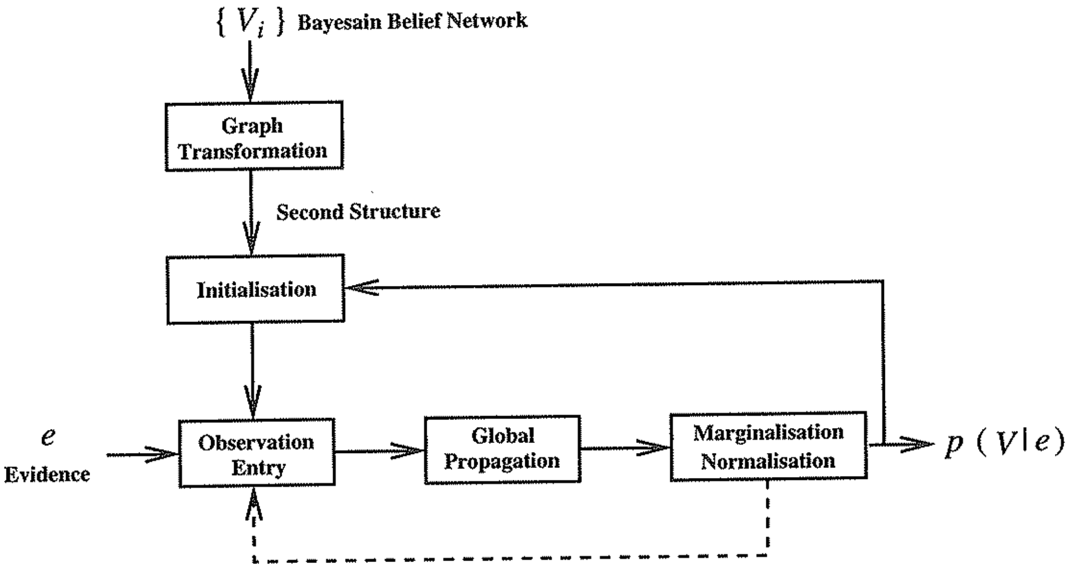


Figure 4.8: *Block diagram of probabilistic inference in a secondary structure.*

After transforming the DAG of a BBN to a secondary structure and initialising the junction tree with potentials representing the joint distribution in Equation (4.9), the structure is ready for inference based on observed evidence. After entering the evidence $\mathbf{e}$ for those clusters with evidence, the potentials $\phi_{\mathbf{C}}$ which represent $P(\mathbf{C})$ (see Equation (4.10)) are modified to contain the evidence and represent $P(\mathbf{C}, \mathbf{e})$. The subsequent probability derivation includes evidence $\mathbf{e}$. Note that if some *evidential variables* are not observed, a BBN can still handle this situation (i.e. incomplete information) by exploiting the built-in causal relations and numerical parameters. After performing global propagation, the potentials of all clusters and separators, $\phi_{\mathbf{C}}$ and $\phi_{\mathbf{S}}$, are modified to $P(\mathbf{C}, \mathbf{e})$ and $P(\mathbf{S}, \mathbf{e})$. Thus, $P(V, \mathbf{e})$ can be obtained from any cluster $\mathbf{C}$ (or separator $\mathbf{S}$) that contains V by marginalisation:

$$P(V, \mathbf{e}) = \sum_{\mathbf{C} \setminus V} \phi_{\mathbf{C}}. \tag{4.13}$$

From $P(V, \mathbf{e})$ of all variables, the posterior $P(V|\mathbf{e})$ can be obtained by normalising $P(V, \mathbf{e})$ as follows:

$$P(V|\mathbf{e}) = \frac{P(V, \mathbf{e})}{P(\mathbf{e})} = \frac{P(V, \mathbf{e})}{\sum_i P(V_i, \mathbf{e})}. \tag{4.14}$$

Given visual evidence $\mathbf{e}$ of different modalities from two cameras which might be uncertain and incomplete, the goal is to find the most likely assignment combination (i.e. a union of m matches between two camera images as defined in Section 4.2.3). Thus, the matching problem defined by Equation (4.1) can now be probabilistically inferred using the secondary structure of a BBN to obtain a probability distribution over the assignment combinations, $A_\alpha = \{A_1, \cdots, A_{m!}\}$. The most likely combination between A_α with the maximum posterior can then be obtained. After marginalisation and normalisation to obtain the probability distribution over the assignment combinations, the network is then re-initialised to incorporate new observations for the next frame.

4.4 A Bayesian Belief Network for Inter-Camera Subject Correspondences

Having introduced the theory of network construction and probability inference in the secondary structure of BBNs, this section describes the design of a BBN for inferring subject correspondences between two camera images. This BBN can effectively fuse multiple visual modalities of different features for matching subjects across cameras. The use of a BBN at a time instant based on a single modality is first described and then the generalisation of the network to fuse multiple modalities over time is explained. Finally, the process of feature validation in both SCT and MCCT for reducing correspondence complexity and defining matching ambiguity is explained.

4.4.1 Feature Correspondence Based on a Single Modality

To infer subject correspondences based on a single modality, the BBN shown in Figure 4.9 is used. As mentioned earlier in this chapter, this method of adaptively fusing multiple modalities based on BBNs was introduced as *Bayesian modality fusion* [171], where the task was to match a single object (i.e. tracking a human head) between image frames from a single camera. This thesis extends this method for matching multiple subjects between image sequences from two widely separated cameras. The nodes of the graph represent the variables of interest. In this network there is one correspondence, one modality confidence, m comparison and k confidence indicator nodes. The true correspondence of multiple subjects between two camera images is represented as a random variable, called the *correspondence* variable, V_c. This unobserved "ground

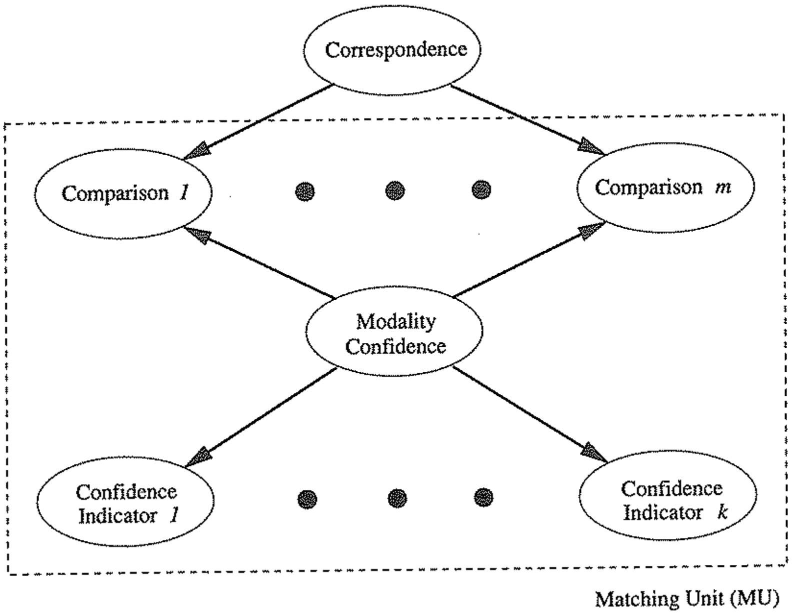


Figure 4.9: *The BBN (Bayesian Belief Network) for inferring the correspondence of subjects between two camera images based on a single modality.*

truth" determines the state of all the *comparison* variables which represent the similarity between the subjects to be matched in two camera images. In order to efficiently fuse all modalities, the confidence of a modality, which represents how reliably the modality reflects the correspondence between two images, is modelled as a *modality confidence* variable. However, this modality confidence variable cannot be directly observed. The evidence regarding the modality observed from the dynamic scene is used to indicate the modality confidence, modelled as *confidence indicator* variables. This unobserved (or hidden) modality confidence variable can be dynamically influenced from the confidence indicator variables. It also influences the comparison variable for adaptively determining the confidence (i.e. weight) of the results of the comparison variables on the subject correspondences. Directed edges with arrows from the unobserved variables (i.e. correspondence and modality confidence) to the evidential variables (i.e. comparison and confidence indicator) capture the probability dependence between these variables. During tracking, the evidential variables are observed from the monitored scene. Then, the observed evidence is set to a discrete state of variable $\mathbf{e}$. This given evidence is then used to infer the probability distribution of the correspondence variable $P(V_c|\mathbf{e})$ for determining the most likely combination of assignments. The four different types of nodes are described as follows:

1. A correspondence node represents a multi-value variable where each value (or state) corresponds to a possible assignment combination $A_\alpha \in \{A_1, \cdots, A_{m!}\}$ where m is the maximum number of subjects in an image. From the computed probability distribution over $\{A_1, \cdots, A_{m!}\}$, given evidence, the correspondence problem defined in Section 4.2.3 is probabilistically inferred. An example of the states represented by the correspondence variable is given in Table 7.3. The prior for correspondence variable is set as all states with equal probability, $\frac{1}{m!}$, for the initial condition (i.e. no observation in the network, more discussed in Section 4.4.2).

2. Comparison nodes: There are m comparison nodes and each node compares one subject in image I_i against all m subjects in image I_j where m is the maximum number of subjects in an image. Thus, all m subjects in I_i are compared to all m subjects in I_j in order to determine the best match. To compare the subjects, the constraints of different modalities for matching are formulated in a statistical framework by defining a similarity measure to quantify the confidence of a possible match. The attribute disparity of the corresponding features is modelled as a Gaussian variable for determining the likelihood of a candidate match (as described in Chapter 5 and Chapter 6, see an example in Figure 5.5 in Section 5.2.2). The experimental results of obtaining Gaussian variable parameters for different modalities are given in Section 7.1. The comparison is based on the MD (Mahalanobis Distance) defined by the Gaussian distribution. The exception is the landmark modality where the matching probability is modelled with uniform distribution (described in Section 5.4.2).

   The comparison nodes are influenced not only by the correspondence variable but also the modality confidence variable, since these variables are represented as parent nodes of comparison nodes. Note that the comparison results in each node allow multiple hypotheses (e.g. two subjects in image I_j are equally similar to a subject in image I_i) which are encoded as the states of a CPT attached to the comparison variable (see an example in Table 7.4). An example of the CPT of a comparison node is given in Table 7.7. The final unique assignment for each subject in I_j to a subject in I_i is determined by using the BBN to fuse multiple modalities and to probabilistically infer the most likely combination of assignments A_α (i.e. a state of correspondence variable).

3. A modality confidence node represents the confidence of the modality in the cor-

respondence dependent on context information. It constrains the influence of this modality on the correspondence variable. This modality confidence variable cannot be directly observed in the image but can be inferred from observed evidence in confidence indicators. An example of the states of a modality confidence variable (i.e. { high confidence, low confidence }) is given in Section 7.3.1. The prior for modality confidence node is set as all states with equal probability for every frame during tracking.

4. Confidence indicator nodes indicate the modality confidence. The notion of confidence indicator nodes is used to reflect that the modality confidence and the estimate of the confidence both vary over time according to the structure of the dynamic scene. To build a coherent framework for adaptively fusing multiple modalities, the unobserved modality confidence is dynamically and probabilistically inferred from visual evidence in the confidence indicator nodes. The visual evidence represented as k nodes in the network enables the BBN model to *context-sensitively* infer the modality confidence. This makes the correspondence adaptively reflect the time-varying confidence of the modality (see an example in Figure 5.6). An example of the states represented by a confidence indicator variable (i.e. { high confidence, medium confidence, low confidence }) is given in Section 7.3.1. An example of the CPT of a confidence indicator node is given in Table 7.8.

In this network, both correspondence node and modality confidence node represent variables to be inferred. From the observed evidence in indicator nodes, the modality confidence is probabilistically inferred. This inferred modality confidence and the computed comparison results are both considered in inferring the probability distribution over the $m!$ assignment combinations. At run-time, the actual number of subjects, n, in both camera images can be less than m. The distribution over $n!$ assignment combinations can be marginalised from the inferred probability distribution over $m!$ assignment combinations. When the numbers of subjects are unequal in the two images (e.g. m subjects in I_i and n subjects in I_j with $m > n$, without loss of generality), the inferred assignment combination has m matches (each match assigns a subject in I_i to a subject in I_j). Thus, each of the m subjects in I_i is assigned one subject in I_j. Those $(m - n)$ subjects in I_i, which are not assigned to any of the n subjects in I_j, are interpreted as not visible in I_j. In other words, only those matches which assign subjects in I_i to one of the n subjects in I_j are valid matches. An example of how to obtain the distribution over the assignment combinations when the subject numbers in either or both camera images are not equal or less than m is given in Section 7.3.1. In order to generalise the BBN for multiple modalities, a Matching Unit (MU) (see Figure 4.9) is defined as the union of all comparison nodes, a modality confidence node and all confidence indicator nodes.

4.4.2 Feature Correspondence Based on Multiple Modalities

After having described the use of the BBN for MCCT at a time instant based on a single modality, the use of a BBN for MCCT over time based on multiple modalities is given here. Figure 4.10 illustrates a BBN for this purpose. The diagram shows how the model fuses n modalities to infer the subject correspondences between two camera images, where each modality is represented with a MU. In the network, there is one correspondence node which represents the subject correspondence between two camera images. The subject correspondences are determined based on n modalities. Each modality has its own modality confidence node. The modality confidence node in each MU defines the relative influence of the comparison results (in the comparison nodes of this MU) on the subject correspondences for this modality. Thus, the modality confidence nodes in different MUs constrain the relative influence on the subject correspondences for different modalities based on the observed evidence in confidence indicator of different modalities.

Figure 4.10 also displays the generalisation of the network to consider the status of variables over time. The representation of temporal dynamics with regard to scene structure in both camera images (e.g. images of subjects and landmarks) provides a temporal pattern of visual evidence for matching subjects across cameras over time. This Bayesian model can capture dependencies between variables at different time instants as well as amongst variables within a time slice. To obtain correspondence consistency, the network is coupled indirectly over time through the specification of prior probability for correspondence node using the posterior of the correspondence inferred in the last frame. As a consequence, the correspondence at each time instant is influenced by the previous matching history. Moreover, to make the matching more reliable and smooth (e.g. less sensitive to noise), the system uses the accumulated evidence in the comparison nodes which compare the subject images to determine the inter-camera subject correspondences. This accumulated evidence is defined as:

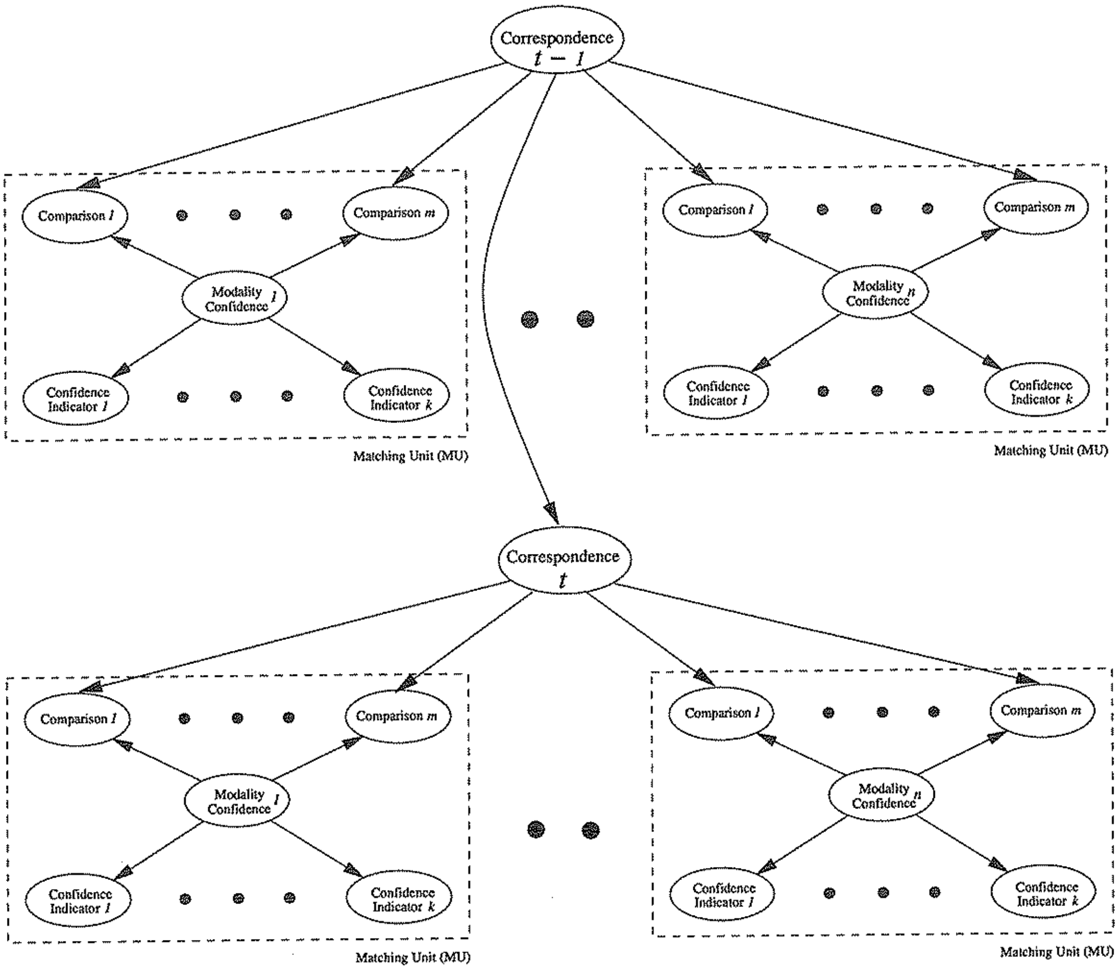


Figure 4.10: *The BBN (Bayesian Belief Network) for inferring the correspondence of subjects between two camera images based on multiple modalities over time.*

$$M = \frac{1}{\sum_{i=0}^{q-1} \alpha_i} \sum_{i=0}^{q-1} \alpha_i \mathcal{F}(l-i), \tag{4.15}$$

where q is the number of frames of the accumulated evidence, l is the frame index, α_i is the weight to set more recent evidence with higher weights and $\mathcal{F}$ is a function estimating the likelihood of the candidate match (discussed in Chapter 5 and Chapter 6). Thus, the visual evidence accumulated over time is integrated into the current network model (see an example in Section 7.3.2). These weights are manually selected in this thesis based on domain knowledge. Note that estimation techniques (e.g. maximum likelihood method) can be used to learn these weights from the experimental data.

Note that if a visual modality becomes less reliable, the comparison can be based on the less reliable evidence in multiple frames due to the accumulated information used for comparison. In such a case, the confidence indicator will lower the influence of the comparison results on the correspondence. On the other hand, if the previous correspondence results are less reliable, the correspondence node can have an unreliable prior. Therefore, the system needs a mechanism to define the matching ambiguity between two camera images in order to avoid using the unreliable correspondence results in the previous frames.

Relationships between Tracking Modes and the BBN

As discussed in Section 3.1, the system first tracks the subjects in each camera (see Figures 4.11) based on its own visual information. In this SCT (Single Camera Tracking) mode, the system matches the subject images (i.e the highest points based on motion vector defined in Section 3.2.2) between successive image frames of a camera over time based on Kalman filtering. To track people using two cameras cooperatively, the system assigns a label (i.e. an identity) to each newly detected subject. If a newly detected subject in an image of a camera I_i has already been tracked and assigned an identity in the other camera image I_j, the system then passes the identity to this subject in I_i by matching subjects across camera images. This MCCT (Multiple Camera Cooperative Tracking) process of matching subjects across camera images can also be used to regain the identity of a subject from other cameras when the identity has been lost and tracking becomes ambiguous in a camera. Moreover, it can be used to check whether different subjects with the same identity in different cameras correspond to the same person.

Figure 4.11 shows the relationships between tracking modes and the BBN used to fuse multiple modalities for MCCT. The system performs SCT in each camera based on its own visual information. The visual information from each of the two cameras is also entered in the BBN at each frame in order to perform MCCT. These different visual modalities are fused by the BBN for determining the inter-camera subject correspondences (i.e. matching each of S_A, S_B and S_C in the left camera image to their corresponding subject, either S_1, S_2 or S_3, in the right). As discussed above, to obtain correspondence consistency, the network is coupled indirectly over time through the specification of prior probability for correspondence node using the posterior of the correspondence inferred in the last frame. Once matching ambiguity is present in a camera, the system then passes subject identities between cameras to resolve the ambiguity in order to track people using two camera cooperatively.

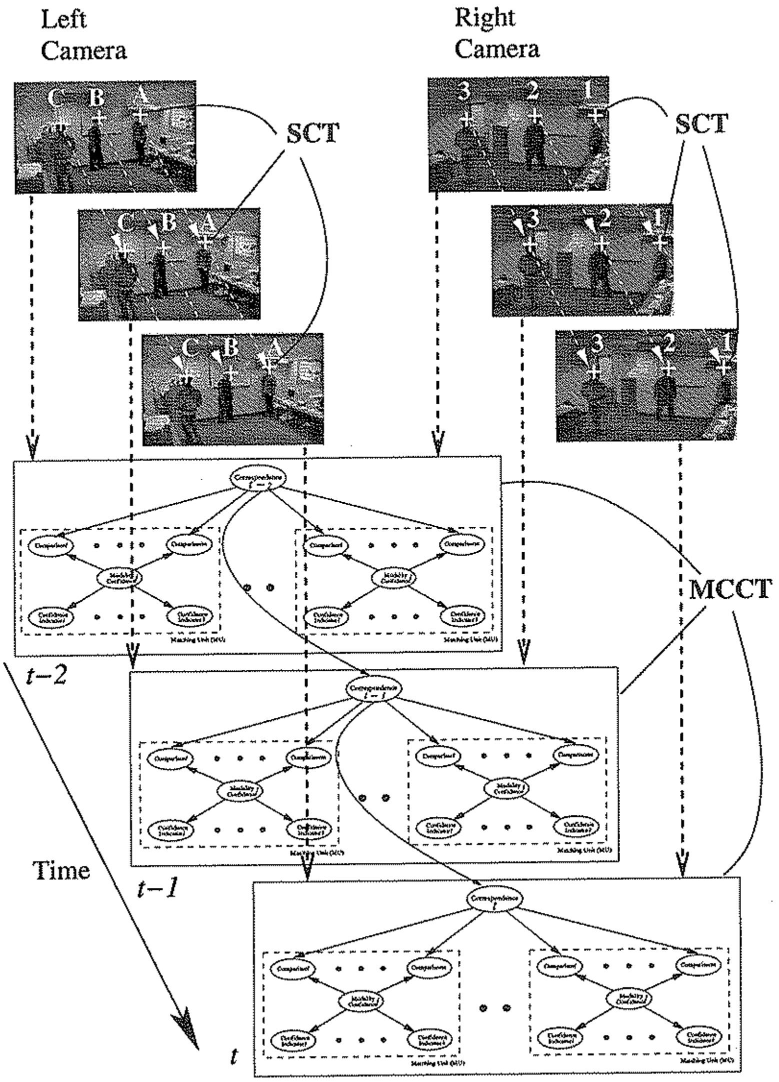


Figure 4.11: *The relationships between tracking modes and the BBN.*

4.4.3 Feature Validation

As discussed above, the system needs to define the matching ambiguity in the inter-camera subject correspondences in order to determine the data collection in the BBN (e.g. stop using the accumulated evidence in the comparison nodes, and the posterior of the correspondence results inferred in the last frame as the prior in the correspondence node). The matching ambiguity in inter-camera subject matching for MCCT and that in SCT (Single Camera Tracking) will be defined in this section. The matching ambiguity is defined in terms of validation of image features. This validation step also relates to the process of handling the complexity problem in correspondence of both MCCT and SCT.

The issue of feature validation is related to the matching process in both SCT and MCCT modes. The features to be considered in SCT are from two successive image frames, $I_i(t)$ and $I_i(t+1)$, of a camera C_i while in the case of MCCT, they are from images $I_i(t)$ and $I_j(t)$ of two cameras C_i and C_j. The goals of this validation process are twofold:

1. to eliminate less likely matches before matching and thus reduce the computational complexity in the correspondence process, and

2. to define the matching ambiguity in order to determine the system operations by testing the matches of the established subject correspondences between two images.

Note that the main difference in the feature validation between SCT and MCCT, as discussed in Chapter 1, is that the features extracted from two cameras for MCCT are from different camera coordinates and different physical processes, thus increasing the matching difficulty and ambiguity. This thesis copes with this problem in MCCT not only by using the BBN to fuse multiple modalities and probabilistically infer the correspondence, but also by explicitly considering the variations of the corresponding features in two camera images (discussed in Chapter 5 and 6). This section focuses on discussing the approach to the complexity problem and defining the matching ambiguities in MCCT as well as those in SCT.

As mentioned earlier in this chapter (Section 4.2.2), the computational complexity problem is an important issue that cannot be ignored in tracking [127]. For the case where there are m subjects in two images, m^m possible assignment combinations need to be considered and the correspondence complexity is exponential. Moreover, the matching in both modes (SCT and MCCT) is based on a comparison of the MD (Mahalanobis Distance). For example, $\mathcal{M}_m$ (i.e. normalised innovation related to the motion vector, see Equation (3.13)) is used in SCT and $\mathcal{M}_h$ (based on homography modality, see Equation (5.7)) is used in MCCT. The computation of the MD, which involves matrix inversion, can also be computationally expensive. Without eliminating some of the less likely matches, the matching process involving computation of MD can slow down significantly.

Validation Gate

Before introducing our methods for feature validation and defining the matching ambiguity, it is useful first to introduce the notion of a *validation gate* [9] (see Figure 4.12), which relates to validation in both SCT and MCCT modes. This validation gate approach places a constraint (shown as a ellipse on which all point are with a given MD in 2D feature space) on the feature value, $\mathcal{M}$ (MD), that candidate matching features might take, for determining valid matches, i.e. only those features for which:

$$\mathcal{M} \leq \mathcal{X}_T^2, \tag{4.16}$$

are valid, where the subscript, $_T$, in $\mathcal{X}_T^2$ is indicative of that it is a threshold, and $\mathcal{X}_T^2$ is a threshold limit based on the $\mathcal{X}^2$ (*chi-square*) statistical test. Figure 4.12 illustrates the validation gate of predicted feature and measured features. The predicted feature is marked with '+', the validated measured features with '●' and the measured features failing the validation test with '△'. This validation of candidate matching features, based on MD, is extensively utilised in robotics and data-association [107]. Since this test can be time-consuming, e.g. matrix inversion, Montiel and Montano [107] proposed to compute MD in a progressive manner. The MD is computed as a non-decreasing quantity. If the MD exceeds $\mathcal{X}_T^2$, the computation stops.

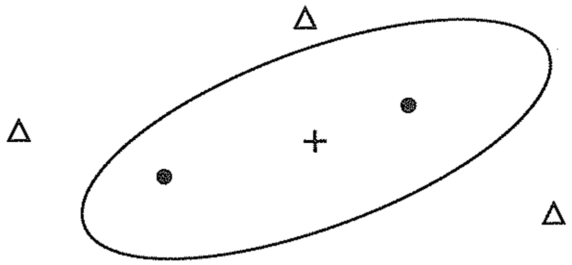


Figure 4.12: *A validation gate.*

This $\mathcal{X}^2$ statistical test leads to the decision of acceptance or rejection of candidate matching features. Since the MD is chi-square distributed with the number of degrees of freedom equal to the dimensionality n_f of the feature vector [108], the probability of the candidate match with respect to the MD between the two features can be obtained from a $\mathcal{X}^2$ distribution table [93]. Thus, one can obtain a value (i.e. $\mathcal{X}_T^2$) corresponding to a pre-defined upper bound of the acceptance MD of the candidate matching features. For example, if the feature vector is one dimensional, $n_f = 1$, and the validation or search range is set so that there is a 95% probability of finding the true feature, the table [93] indicates that the corresponding MD is 3.84. So $\mathcal{X}_T^2$ is set to 3.84 (see an example in Figure 7.7 in Section 7.1.4). Conversely, if a feature fails the inequality test of Equation (4.16), there is a chance $\leq 5\%$ that this feature is the correct corresponding feature.

Feature Validation for SCT

There are some methods available to handle the complexity problem in SCT [127], such as small velocity change, smooth motion constraints [64] and the bucket method [190]. Zhang and Faugeras [190] partitioned the search image space into *buckets* (square blocks, see Figure 4.13). A disc is defined by setting its centre to coincide with the predicted feature point on the image plane. Only those feature points located in the buckets which the disc intersects are valid feature points. Figure 4.13 shows that S_1 and S_4 are the valid feature points which are in the buckets intersected by the disc, and S_2 and S_3 are not valid. This bucket method is similar to the validation gate method discussed above, but constraint is placed on the image distance instead of the MD in the feature space.

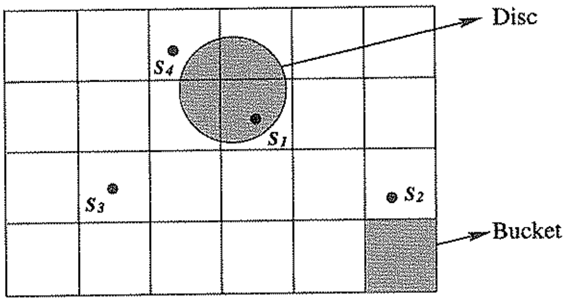


Figure 4.13: *The bucket method.*

The bucket method is adopted in this thesis for feature validation in SCT. This method is used to compare the predicted highest point of a subject from the previous frame and the extracted highest points of subjects in the current frame. Only those highest points inside the disc on the image are valid. Only for these valid features, the system needs to further compute the $\mathcal{M}_m$ (see Equation (3.13)). By computing the $\mathcal{M}_m$ for different highest points, the system can determine which subject is to be associated to a Kalman filter in order to keep tracking the subject. Note that the predicted highest point (x, y) of a subject is obtained from the predicted motion vectors $(\hat{\mathbf{z}}_x(k+1|k) = [x, \dot{x}, \ddot{x}]^T$ and $\hat{\mathbf{z}}_y(k+1|k) = [y, \dot{y}, \ddot{y}]^T$, (defined in Section 3.2.2). These are predicted based on Kalman filtering (described in Section 3.2.2).

In applying the bucket method for feature validation, there could be no features inside the disk, e.g. as a result of a sudden change in motion. In such cases, the acceptance region is enlarged. On the other hand, when more than one candidate matching feature falls inside the validation region, a decision must be made as to which of the features in I_j is to be assigned the feature in I_i. In SCT, the simple nearest neighbour method (described in Section 3.2.2) is used to compare the related $\mathcal{M}_m$ of different subjects.

In order to detect a questionable match during SCT (as discussed in Section 2.2.1), the $\mathcal{X}^2$ test is used to test the $\mathcal{M}_m$ (see Equation (3.13)) of the matched corresponding subjects in two contiguous image frames, $I_i(t)$ and $I_i(t+1)$. In this case, $n_f = 3$ (i.e. dimensionality of the motion vector defined in Section 3.2.2) and the validation is set so that there is a 95% probability of finding the true feature, the table [93] indicates that the corresponding MD is 7.81. So $\mathcal{X}_T^2$ is set to 7.81. Once a match between two corresponding subjects fails the test, the system performs MCCT to obtain the correct identities from the other camera in order to resolve the matching ambiguity. Moreover, the system can also detect a questionable match in SCT by performing MCCT to test if the matching in one camera is compatible with the other.

Feature Validation for MCCT

In general, methods for reducing complexity in SCT cannot be applied to MCCT using widely separated cameras. In SCT, constraints (e.g. velocity change, smooth motion constraints) are used to reduce the number of candidate matches between consecutive frames where the image variations are assumed to be relatively small. In MCCT, the two camera images $I_i(t)$ and $I_j(t)$ are obtained using different camera coordinates. To apply these constraints, the image positions of objects in two camera images need to be transformed to a common coordinate system. Moreover, the image variations between the two widely separated cameras generally cannot be ignored.

To handle the complexity problem in MCCT, it is necessary to use some global constraints. One possibility is to apply homography as a global constraint (see Section 5.2) and use the ground plane constraint (see Section 2.2.2) which assumes the lowest parts of the subjects to be matched are visible in both images. This validation can be done by comparing the estimated image position of a subject's lowest point in I_j with the observed lowest points of the subjects in I_j. Thus, by applying the bucked method, only those subjects' lowest points located in the buckets which the disc (its centre is set as the estimated subject's lowest point) intersects are valid subjects. However, in a cluttered environment, the lowest points of subjects may not be visible. Therefore, in this thesis, the validation gate method is performed by using homography related to the virtual plane of a subject's highest points. To reduce the complexity, the less likely matches are eliminated by testing their $\mathcal{M}_h$ (defined in Section 5.2.2). Only those subjects with $\mathcal{M}_h$ smaller than the threshold value $\mathcal{X}_T^2$ are valid.

In each comparison node, m subjects in I_j need to be compared to one subject in I_i. To avoid computing the MD for all m subjects for all modalities, only those subjects in I_j which are validated by the homography modality need have their MD computed further. This is because in comparison with other modalities, homography is a more powerful constraint which can ideally find a corresponding point between two camera images [59]. Thus, it is used as a global constraint in our inter-camera correspondence problem. In applying the validation gate method based on homography, similar to SCT, when there are not any subjects inside the gate, the acceptance region is enlarged. On the other hand, when more than one candidate matching feature falls inside the validation region based on the homography modality, all are compared in the comparison nodes of other modalities. Note that, as mentioned earlier, the comparison results of the subjects in each comparison node allow multiple hypotheses to be made (see an example in Table 7.4). The unique correspondence of each subject in two images is probabilistically inferred in the BBN. Another method for reducing the number of candidate matches is to use domain knowledge such as the landmark method [28] (see Chapter 5) or the spatial relationship between FOVs of different cameras to constrain the matching [75]. This knowledge-based method can constrain the image positions of corresponding subjects in both camera images and serve as a validation tool.

To define the matching ambiguity in MCCT, the $\mathcal{X}^2$ test is used to test each match in the assignment combination obtained in the previous frame. If any match has more than one modality larger than the threshold, $\mathcal{X}_T^2$ (see an example in Figure 7.7 in Section 7.1.4), the system does not use previous matching results as a prior in the correspondence node. Moreover, the number of frames of accumulated evidence used in the comparison node is set as $q = 0$ for all modalities to prevent using less reliable evidence. Thus, Equation (4.15) becomes $\mathcal{F}(l)$ and the system compares subjects based on the information in the current frame images.

4.5 Summary

This chapter has developed a framework based on the BBN for integrating multiple modalities over time in order to match subjects across camera images when performing MCCT. The complexity problem in establishing correspondence for both SCT and MCCT is handled by applying some global constraints. The questionable match in both SCT and MCCT is defined and used during tracking for making the matching more reliable. The theory for inferring the subject correspondences in the BBN has been given in Section 4.3.4. The BBN (see Figure 4.10) used for fusing multiple modalities in order to infer the subject correspondences has been described in Section 4.4.2. To infer the subject correspondences using the BBN, each modality, used as a constraint on the inter-camera subject correspondences, is constructed with a single MU. Chapters 5 and 6 describe the details of the data collection in the evidential nodes (i.e. comparison and confidence indicator) in the BBN for different modalities (see Table 4.2). The collected evidence from different modalities is then used to probabilistically infer the subject correspondences between two camera images. The next chapter introduces the geometry-based modalities.

Table 4.2: *Modalities for inter-camera feature correspondence.*

Geometry-Based Modalities (Chapter 5)	Recognition-Based Modalities (Chapter 6)
homography	apparent colour
epipolar	apparent height
landmark	

Chapter 5

Geometry-Based Modalities

5.1 Introduction

The previous chapter described the use of a BBN to adaptively fuse multiple visual modalities for matching subjects across camera images over time in order to perform MCCT. This chapter describes the geometry-based modalities used in the BBN as local constraints for matching subjects across cameras in MCCT and explains the details of the data collection in the evidential nodes of the BBN (see Figure 4.10). Note that each modality corresponds to a MU in the BBN for inferring the subject correspondences by fusing multiple modalities. Since the subject features used for determining inter-camera subject correspondences are obtained from two widely separated cameras with different camera coordinates, the image variations can be significant. Direct use of the image coordinates for feature correspondence between two camera images does not always make a correct match (see Figure 5.3). The main purpose of using geometric modalities is to handle this problem by finding the geometric positional relationships between the corresponding subjects in two camera images. With these geometric constraints, the search space can be reduced so that the feature correspondence can be established more reliably.

The work presented in this chapter uses multi-view geometry (including homography and epipolar geometry) and landmarks for solving inter-camera subject correspondences. The novelties are twofold. The first lies in the use of multi-view geometry with an explicit consideration that in a cluttered indoor environment, the lower part of a person can be invisible (see Figure 6.3). This thesis proposes to use the highest point of the subject image to overcomes this problem [27]. The second is the use of knowledge-based natural landmarks in the scene to reason about the spatial positional relationships of the corresponding subjects in two camera images [28].

Geometry and Computer Vision

Computer vision is concerned with the development of machines that can automatically analyse and interpret the images of scenes. Most problems in computer vision can be couched in geometric terms, and geometric methods constitute one of the most useful tools for this analysis [170]. Though Euclidean geometry describes our 3D world well, it is insufficient in the context of the imaging process of a camera (e.g. lengths and angles are no longer preserved, and parallel lines may intersect). Projective geometry deals elegantly with the general case of perspective projection and therefore provides understanding of the geometric aspect of image formation and 3D vision [106, 48] , e.g. it can deal with projections and objects at infinity, though it lacks the notations of angles or distances (due to distortion). The adoption of projective geometry, as a supplement to Euclidean geometry, provides a useful approach to many computer vision problems and has led to improved recognition methods and a better understanding of the geometry of multiple views, particularly in the case of un-calibrated cameras. This chapter uses some modalities based on multi-view geometry which do not require camera calibration. A good introduction to the subject of multi-view geometry can be found in the book by Hartley and Zisserman [59]. Note that one major characteristic of computer vision problems, in contrast with projective geometry as pure mathematics, is that data is not necessarily accurate [84]. For example, the positional ambiguity of the highest point that is used for representing subject position can degrade the matching reliability. Therefore, a combination of geometry and statistics (i.e. a BBN) is adopted in this thesis in order to handle data uncertainty for making inter-camera subject correspondences more robust.

Multi-View Geometry

Multi-view geometry involves analysis of the geometric relations that exist between the images of objects from different views. These relations are important to be understood not only for providing explanations of appearances in different views, but also because their understanding is important for a range of applications [59], such as self calibration [7] (computing the intrinsic camera parameters using only information in the images), scene reconstruction [165] (reconstructing the 3D structure of the objects in the scene) and structure from motion [67] (analysis of image motion caused by relative motion between objects and cameras). One of the advantages of applying multi-view geometry is that some of the relationships between multiple views can be obtained from un-calibrated cameras. For example, both homography and epipolar geometry relate to the geometry between two views, and trifocal tensor relates to the geometry between three views [169]. The advantage of multi-view geometry is that it avoids computation of camera parameters and hence results in a simpler algorithm [59].

Scene Knowledge

The other geometric modality used in this thesis is based on the knowledge of landmarks in the scene. Scene knowledge uses spatial context information about the scene structure for image understanding and interpretation [150]. This high-level knowledge can be used for various problems in computer vision. Scene knowledge is often referred to as a *model*. For example, the knowledge of an object model can be used to interpret the contours of the object image. The landmark modality used for inferring the spatial relationships of the corresponding subjects in two cameras images is a type of 3D *world model*. A world model is generally used to store information concerning the state of the environment in which the system is operating [98]. Depending on the needs of particular applications, information stored in the world model can take many different formats. High level reasoning processes can incorporate the world model to make inferences in order to direct the processing and/or operation of the system, such as to navigate a robot [143] or to guide a tracking system to follow people in a building [140]. Our method [28] uses the world model to geometrically reason about the subject correspondences between two un-calibrated camera images and obtain the relative positions of people in the scene. This type of *geometry reasoning* method is a knowledge-based vision technique combining domain knowledge and image processing [180]

The Approach

The problem that this chapter aims to solve is that given a point $\mathbf{p}$ in the first camera image, I_1, how does one use geometric methods to constrain its corresponding point $\mathbf{p}'$ in the second camera image I_2. This question will be answered for three cases as follows:

1. For images of points on a common plane in the 3D world, the corresponding point is uniquely determined by applying homography (Section 5.2).

2. For arbitrary image points, the corresponding point is constrained to lie on a line (an epipolar line) by applying epipolar geometry (Section 5.3).

3. For images of points viewed in front of landmarks, the corresponding point is constrained to lie in an limited area by using the knowledge based on the vertical line landmarks (Section 5.4).

In all three cases, constraints can be computed from prior knowledge of image correspondences, and do not require camera calibration (i.e. the process to estimate camera parameters, an introduction to camera calibration is given in Appendix A). The first two cases need prior knowledge of point correspondences and the third needs the knowledge of correspondence of the line landmarks between the two camera images. To reduce the computational cost, a single feature point (the highest point) is extracted from each subject image and used for representing the subject's image position. The preprocessing step of feature extraction of the highest point is described in Section 3.2.1 (see Figure 3.7). Note that the prior knowledge of the correspondence required for these modalities only needs to be obtained once off-line, although it needs to be re-established if the cameras are moved.

5.2 Homography Modality

This section explores homography as a constraint on point correspondence between two camera images. Given a set of corresponding points in two images which lie on a scene plane in the world, the correspondence of image points on this scene plane can be uniquely determined by applying homography. This section first introduces the theory of homography regarding a scene plane and estimation of homography, then describes how to apply homography to the subject correspondence problem.

5.2.1 A Scene Plane and Homography

As mentioned above, if a scene point $\mathbf{P}$ lies on a plane (see Figure 5.1), then the image point $\mathbf{p}$ in the first image I_1, corresponding to $\mathbf{P}$, determines the image position of $\mathbf{p}'$ in the second image I_2, which also corresponds to $\mathbf{P}$. This planar projective transformation is called *homography* and the homography between two images is said to be *induced* by the scene plane [59]. The image points relationship is expressed as $\mathbf{p}' = \mathbf{H}\mathbf{p}$ where $\mathbf{H}$ is the 3×3 homography matrix. Figure 5.1 illustrates the concept of homography. The ray corresponding to a point $\mathbf{p}$ in the first image I_1 is extended to meet the scene plane at a point $\mathbf{P}$. This point is projected to a point $\mathbf{p}'$ in the second image I_2. The projective

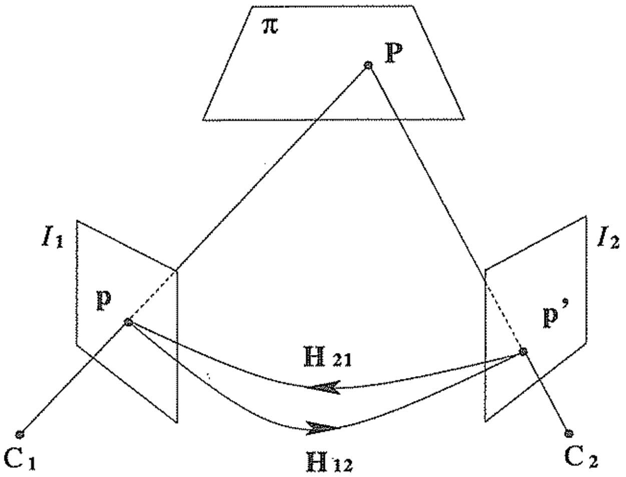


Figure 5.1: *Illustration of homography induced by a scene plane π between two camera images with camera centres C_1 and C_2. The camera centre is defined as the centre of projection.*

transformation between two images induced by a scene plane is called homography, $\mathbf{H}$. The homography $\mathbf{H}_{12}$ transfers points from I_1 to I_2 while $\mathbf{H}_{21}$ transfers points from I_2 to I_1, with $\mathbf{H}_{12} = \mathbf{H}_{21}^{-1}$. The derivation of homography is described as follows.

In homogeneous coordinates, a point is represented as $\mathbf{p} = (x_1, x_2, x_3)^{\mathrm{T}}$ which corresponds to the point $(x_1/x_3, x_2/x_3)$ in the image coordinates; $\mathbf{p}'$ and $\mathbf{Hp}$ have the same direction but may differ by a non-zero scale factor [59]. The equation may be expressed in terms of a vector cross product as $\mathbf{p}'_i \times \mathbf{Hp}_i = 0$ for a set of given corresponding points, $\mathbf{p}_i \leftrightarrow \mathbf{p}'_i$, in two images I_1 and I_2. By denoting the jth row of the matrix $\mathbf{H}$ as $\mathbf{h}^{j\mathrm{T}}$, then one may write

$$\mathbf{Hp}_i = \begin{bmatrix} \mathbf{h}^{1\mathrm{T}}\mathbf{p}_i \\ \mathbf{h}^{2\mathrm{T}}\mathbf{p}_i \\ \mathbf{h}^{3\mathrm{T}}\mathbf{p}_i \end{bmatrix}, \tag{5.1}$$

where

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \mathbf{h}_3 \\ \mathbf{h}_4 & \mathbf{h}_5 & \mathbf{h}_6 \\ \mathbf{h}_7 & \mathbf{h}_8 & \mathbf{h}_9 \end{bmatrix} = \begin{bmatrix} \mathbf{h}^{1\mathrm{T}} \\ \mathbf{h}^{2\mathrm{T}} \\ \mathbf{h}^{3\mathrm{T}} \end{bmatrix}. \tag{5.2}$$

With $\mathbf{p}' = (x'_1, x'_2, x'_3)^T$, the cross product may then be written as

$$\mathbf{p}'_i \times \mathbf{H}\mathbf{p}_i = \begin{bmatrix} x'_2\mathbf{h}^{3\mathrm{T}}\mathbf{p}_i - x'_3\mathbf{h}^{2\mathrm{T}}\mathbf{p}_i \\ x'_3\mathbf{h}^{1\mathrm{T}}\mathbf{p}_i - x'_1\mathbf{h}^{3\mathrm{T}}\mathbf{p}_i \\ x'_1\mathbf{h}^{2\mathrm{T}}\mathbf{p}_i - x'_2\mathbf{h}^{1\mathrm{T}}\mathbf{p}_i \end{bmatrix}. \tag{5.3}$$

Because $\mathbf{p}'_i \times \mathbf{H}\mathbf{p}_i = 0$ and $\mathbf{h}^{j\mathrm{T}}\mathbf{p}_i = \mathbf{p}_i^{\mathrm{T}}\mathbf{h}^j$, Equation (5.3) can be re-written as

$$\begin{bmatrix} 0^{\mathrm{T}} & -x'_3\mathbf{p}_i^{\mathrm{T}} & x'_2\mathbf{p}_i^{\mathrm{T}} \\ x'_3\mathbf{p}_i^{\mathrm{T}} & 0^{\mathrm{T}} & -x'_1\mathbf{p}_i^{\mathrm{T}} \\ -x'_2\mathbf{p}_i^{\mathrm{T}} & x'_1\mathbf{p}_i^{\mathrm{T}} & 0^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{h}^1 \\ \mathbf{h}^2 \\ \mathbf{h}^3 \end{bmatrix} = 0. \tag{5.4}$$

Equation (5.4) corresponds to a linear system where the unknown $\{\mathbf{h}^j\}$ are the row vectors of the homography matrix $\mathbf{H}$. Although there are three equations in Equation (5.4), only two of them are linearly independent. It is usual to omit the third equation. The set of equations in Equation (5.4) then becomes:

$$\begin{bmatrix} 0^{\mathrm{T}} & -x'_3\mathbf{p}_i^{\mathrm{T}} & x'_2\mathbf{p}_i^{\mathrm{T}} \\ x'_3\mathbf{p}_i^{\mathrm{T}} & 0^{\mathrm{T}} & -x'_1\mathbf{p}_i^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{h}^1 \\ \mathbf{h}^2 \\ \mathbf{h}^3 \end{bmatrix} = 0. \tag{5.5}$$

Each point correspondence gives rise to two independent equations in Equation (5.5). Given 4 point correspondences, $\mathbf{H}$ can be determined up to a scale factor. However, since the points may not be correctly extracted from the image due to noise or system error, more corresponding points can be used to estimate $\mathbf{H}$ more accurately. If more than four point correspondences are given, then the set of equations from Equation (5.5) is over-determined. The Singular Value Decomposition (SVD) method [124] is used to obtain a better estimation of $\mathbf{H}$ with more than 4 corresponding points. There are also some optimisation methods which can be used for estimating $\mathbf{H}$ by minimising a cost function based on image distance [59]. On the other hand, homography can also be obtained from the geometric relationship between the scene plane and the relative pose (position and orientation) of two cameras if the two cameras are calibrated [59].

Homography is used in a range of computer vision problems, e.g. motion estimation [173] and stereo matching [125]. However, this relation holds only when the points in two camera images lie on the same scene plane, or the two images are from a rotated camera [167]. Conversely, if the scene plane contains one (or both) of the camera centres, then the homography degenerates [59]. For example, if the scene plane contains the second camera centre C_2, all points in the first image which are on the scene plane are mapped to points on a line (where the scene plane intersects the image plane) in the second image I_2.

5.2.2 Using Homography for Inter-Camera Subject Correspondences

In general, when homography is applied to the problem of matching subjects across different cameras, the ground plane constraint (see Section 2.2.2) is used (assuming the lowest points are visible in the images of all cameras [92]). However, as previously mentioned, the lower parts of subjects may not always be visible. Instead, the highest point of a subject is used here for applying homography, with the assumption that the highest points lie on a *virtual plane*, which is parallel to the ground plane, as a person is moving (see Figure 5.2). The map from p to p' is the homography induced by the virtual plane π which contains the highest point P of the person.

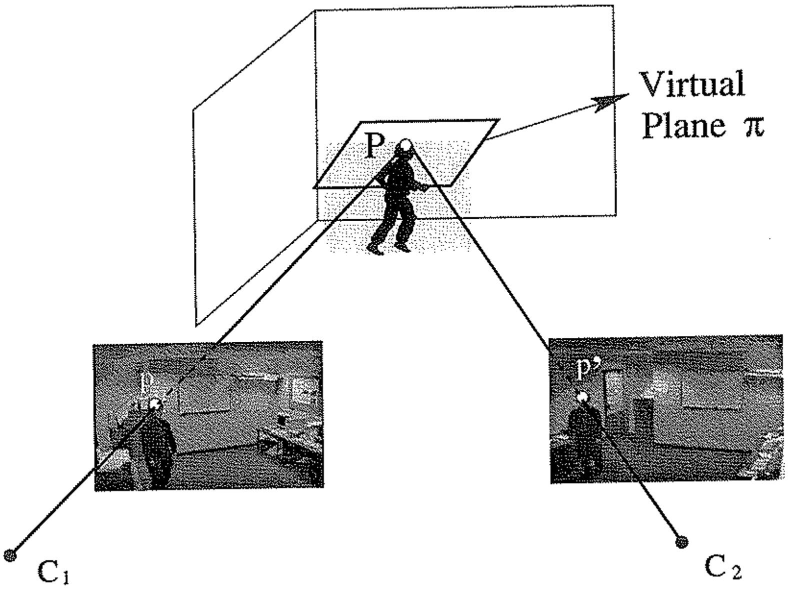


Figure 5.2: *The homography induced by the virtual plane of a person's highest point.*

To establish the homography for the virtual plane containing a person's highest point, the system has to obtain the point correspondences on the virtual plane between two camera images. The point correspondences are obtained from the extracted highest points of the person's images in two cameras. However, before the system matches a subject in two cameras, the point correspondences on the virtual plane are not available. The system has to match subjects based on other modalities first in order to obtain the corresponding highest points in two camera images. Thus, for different people with different heights, the system estimates different homographies induced by different virtual planes of these people. Once the identity of a subject is lost in I_2, but people are tracked with identities in I_1. The system can track subjects with identities in I_2 after occlusion by passing subject identities from I_1 to I_2. Thus, the system needs to match subjects in two camera images in order to pass subject identities. From the established $\mathbf{H}_{12}$ for people tracked in I_1, the highest points $\mathbf{p}$ of subjects in I_1 are transformed to I_2 as $\mathbf{p}\prime = \mathbf{H}_{12}\mathbf{p}$. The transformed points in I_2 can be used for matching subjects to regain the identity of the subject in I_2. Although this modality might be less reliable when a person changes attitude (e.g. kneels down) dramatically such that the highest point does not lie on the virtual plane, it works well for matching walking people. To handle this problem, the node of confidence indicator in the BBN (see section 4.4.1) is used to adaptively reduce the modality confidence (explained later in this section) in order to reduce the relative influence of the homography modality on the subject correspondences between two camera images.

Figure 5.3 illustrates an example where homography is necessary to estimate the subject position across camera images. Two subjects are seen in both camera images. The image positions of subjects are represented by their extracted highest points marked with '+'. In searching for the subject corresponding to subject A (S_A) in the left image, the established homography induced by the virtual plane containing his highest point (see Figure 5.2) is used to transfer S_A's highest point '+' to '$\diamond$'. This point, '$\diamond$', can be used to correctly match S_2 to S_A, based on image distance. One cannot use the same (x, y) image position of the feature point (without transformation) to search for the corresponding subject across camera images. For example, directly applying the coordinates of S_A's highest point '+' to the left image (marked with '$\circ$') in order to search for the corresponding subject results in an incorrect match: S_1 is matched to S_A. Without this homography transformation, the direct use of the image position in the other camera image for matching can result in a incorrect match due to large image variations between two images from widely separated cameras.

Figure 5.4 shows an example of the transferred highest point across camera images by homography for 40 frames. The observed highest point (white cross) of the subject in the right camera image is transferred to the left camera image (white cross) based on the on-line established homography. The observed and transferred trajectories of these two feature points over 40 frames are plotted in both camera images. One limitation

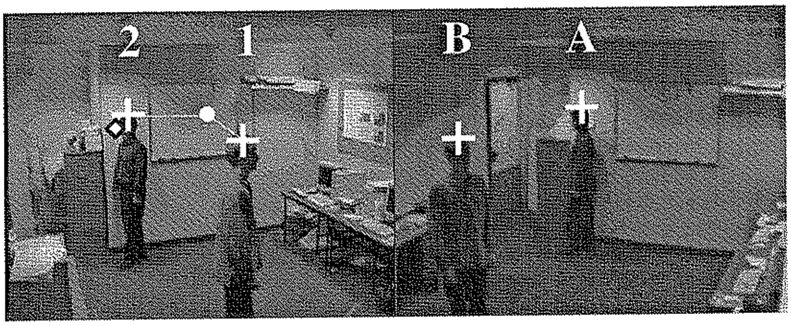


Figure 5.3: *An example of a case where homography is necessary for inter-camera subject correspondences.*

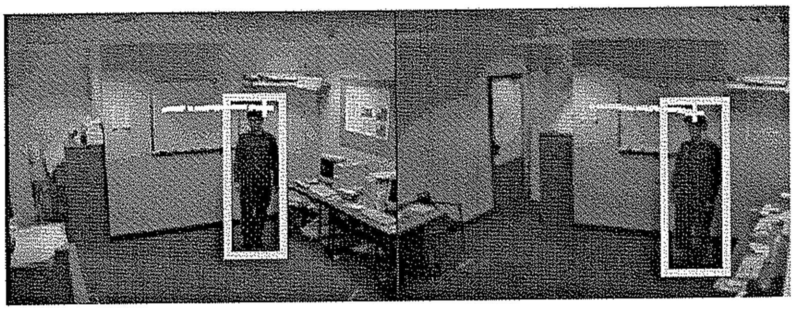


Figure 5.4: *An example of applying the homography induced by the virtual plane (containing the person's highest points) to transfer points across cameras.*

of this method is that the position of the camera must be high enough such that the homography does not degenerate as the virtual plane projected as a line on the image. This is because the homography induced by a virtual plane can be used to transfer points, on this virtual plane, between two camera images. When the camera is not high enough, the virtual plane can be projected as a line, instead of a region, in the camera image. Thus, all the transferred points in a camera image lie on this line and the positions of these points may not be far from each other to make a correct correspondence.

Kinematic Vector for Matching

To match the subjects in two camera images, I_1 and I_2, the highest point $p(x, y)$ of a subject in I_1 is first transferred to a point $p'(x', y')$ in I_2. The transferred points are used to compute $\mathbf{x}' = (x', y', \dot{x}', \dot{y}')$ (called the *kinematic vector*) for searching for the corresponding subjects in I_2, where $(\dot{x}', \dot{y}')$ is the spatial displacement of the transferred point between consecutive frames. This kinematic vector is similar to the motion trajectory [25], where speed, direction and curvature of the trajectory are used to identify different motion events, e.g. a sudden change of direction or stopping. Since it is assumed here that the homography is correctly estimated, the inter-camera transferred highest point ideally coincides with the observed highest point of the corresponding subject. Thus, the likelihood of a subject S_a in I_2 being the subject S_1 in I_1 can be computed from the disparity between the estimated kinematic vector $\mathbf{x}'$ (computed from the highest point of S_1) and the observed kinematic vector $\mathbf{x}$ of S_a. The likelihood of a candidate matching subject should be a decreasing function of this disparity value. Such attribute disparity, $\triangle\mathbf{x} = (\mathbf{x} - \mathbf{x}')$, is assumed to be a Gaussian distribution with a zero mean. Thus, the matching likelihood of a subject with observed kinematic vector $\mathbf{x}$ in I_2 is given by the probability density function:

$$f(\triangle\mathbf{x}) = \frac{1}{2\pi|\mathbf{\Sigma}|^{1/2}} exp(-\frac{1}{2}[(\triangle\mathbf{x})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\triangle\mathbf{x})]). \tag{5.6}$$

Figure 5.5 shows an example of using a 1D Gaussian variable modelling the matching likelihood. Since disparity $| \triangle x_2| < | \triangle x_1|$, the matching candidate with $\triangle x_2$ is more likely to be the correct match than the other candidate with $\triangle x_1$.

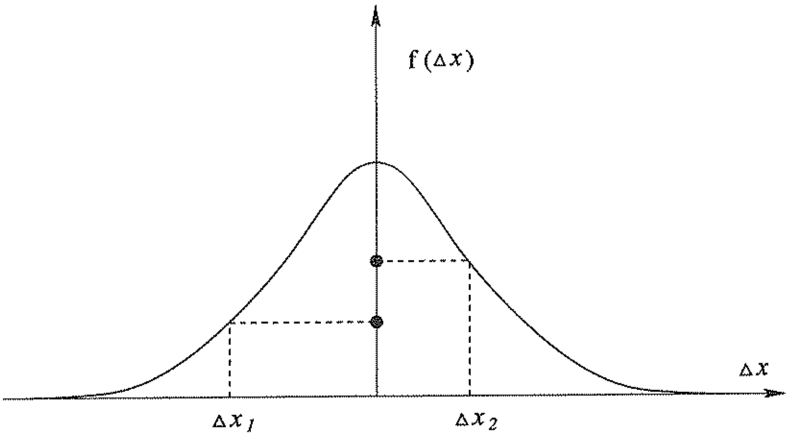


Figure 5.5: *An example of applying a 1D Gaussian variable with zero mean to model the matching likelihood given the attribute disparity, $\triangle x$.*

To compare the candidate matching points of different subjects based on homography,

$$\mathcal{M}_h = [(\triangle\mathbf{x})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\triangle\mathbf{x})] \tag{5.7}$$

is used. This MD (Mahalanobis Distance), $\mathcal{M}_h$, is used for comparing the subjects in the comparison nodes in the BBN (Figure 4.10). The distribution of the distance is learnt from a set of examples (see an example in Section 7.1.1). Since the accumulated information is used for the comparison of subjects in the comparison node (as discussed in Section 4.4.2, see Equation (4.15)), comparison of the homography modality is based on:

$$M = \frac{1}{\sum_{i=0}^{q-1} \alpha_i} \sum_{i=0}^{q-1} \alpha_i \mathcal{M}_h(l-i), \tag{5.8}$$

where q is the number of frames of accumulated evidences, l is the frame index and α_i is the weight used for setting more recent evidence with higher weights. Thus, the comparison is based on the accumulated kinematic vectors of q frames (i.e. a trajectory) which can be described by a list of motions, including position and velocity information [94]:

$$[(x, y, \dot{x}', \dot{y}')_l, \cdots, (x, y, \dot{x}', \dot{y}')_{l-(q-1)}]. \tag{5.9}$$

To define the comparison results of two subject based on homography modality, the $\mathcal{X}^2$ (*chi-square*) statistical test is used again. As mentioned in Section 4.4.3, the MD is chi-square distributed with the number of degrees of freedom equal to the dimensionality n_f of the feature vector [108]. In this case, $n_f = 4$ for Kinematic vector and $\mathcal{X}_T^2$ corresponding to a 95% probability of finding the true feature is 9.49 [93]. So $\mathcal{X}_T^2$ is set to 9.49. In the comparison node, the comparison result between each pair of subjects is defined such that it has two relationships which are *similar* and *not similar* (see an example in Table 7.4). The comparison result is determined by:

$$C_p(l) = \begin{cases} \text{similar} & , \quad M \leq \mathcal{X}_T^2. \\ \text{not similar} & , \quad M > \mathcal{X}_T^2. \end{cases} \tag{5.10}$$

where l is the frame index. Note that the results in comparison nodes in different MUs of different modalities have the same two relationships. These comparison relationships for all modalities are determined by the same methods (i.e. Equation (5.10)) with the exception of landmark modality.

Although there are general algorithms for continuous variables (e.g. Gaussian variables or non-parametric density variable [79]), there are some constraints on the architecture. For example, a discrete node cannot be a child of a continuous node. Also, the conditional probability function for a continuous node (a regression-style function) may not always be appropriate. Therefore, this thesis adopts discrete random variables in the BBN and all observations are discretised.

After describing data collection in the comparison node, two confidence indicators, which indicate the modality confidence on the inter-camera correspondence, are defined for the homography modality in the BBN (Figure 4.10). Experiments performed show that the reliability of the homography modality degrades when the feature points are not extracted accurately. When the highest points vary significantly between consecutive frames, it may be the result of image noise or a sudden change in subject pose, and so the modality reliability might be lower (see Figure 5.6). This is because in this case the established homography might be less reliable in transferring points. To indicate the significance of the variation in the highest point, the *segmentation status* of the highest point of subjects is used and defined as follows. Figure 5.6 shows two consecutive frames from two cameras. The top row is the 50^{th} frame and the bottom the 51^{st}. The right-hand side graph in the bottom row shows the related image distance in the right camera image used to indicate modality confidence (discussed below). Two subjects are seen in both camera images. The image positions of subjects are represented by their highest points '+'. To search for the subject corresponding to subject A (S_A) in the left camera image, the homography (see Figure 5.2) is used to transfer '+' to '○'. The transferred points '○' can correctly determine the correspondence for the 50^{th} frame but can be less reliable for 51^{st} due to an incorrect extracted highest point. In the 51^{st} frame, the transferred point '○' is further from the highest point of S_1 than in the case where the highest point of S_A is correctly extracted. In order to dynamically adjust the modality confidence of the correspondence, the distance, d_c (see the right-hand side graph in the bottom row in Figure 5.6), between the highest points of a tracked subject in two consecutive frames is used. For S_A, the d_c between the extracted highest point '+' in the 51^{st} frame and the extracted highest point in the 50^{th} frame (which is also shown as '$\diamondsuit$' in the 51^{st} frame) is significant. Thus, the modality confidence is reduced. The image distance between the highest points (x, y) of a tracked subject in two consecutive frames is used to indicate modality confidence and defined as:

$$d_c = \sqrt{(x(l) - x(l-1))^2 + (y(l) - y(l-1))^2}, \tag{5.11}$$

where l is the frame index. The segmentation status of the highest point is defined as the mean distance, $D = \frac{1}{m}\sum_{j=1}^{m} d_{c,j}$ (pixels), of m subjects' highest points in a camera image. This value D is used as a confidence indicator to reduce the confidence when this distance is large. The state of this confidence indicator is defined as:

$$C_i(l) = \begin{cases} \text{high confidence} & , \quad D \leq 2.0 \\ \text{medium confidence} & , \quad 2.0 < D \leq 4.0 \\ \text{low confidence} & , \quad D > 4.0 \end{cases} \tag{5.12}$$

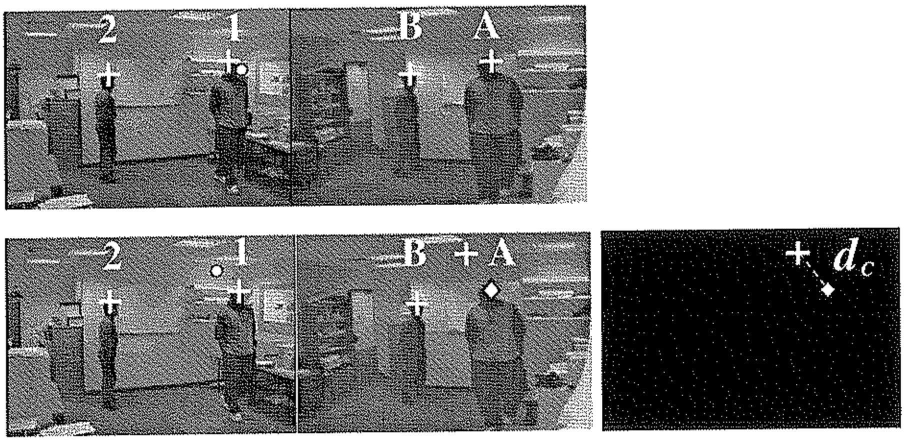


Figure 5.6: *An example of the confidence indicator.*

Note that the range for determining the state of confidence indicator is based on domain knowledge and may only suitable for the camera setup used in this thesis. Moreover, experiments performed also show that when the image positions of subjects are close, the reliability of homography degrades. Therefore, the mean distance of the highest points between all m subjects in a camera image, defined as $D = \frac{1}{a}\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m} d_{i,j}$ (pixels), where $a = \sum_{i=1}^{m-1} i$ and d is the distance between the highest points of two subjects in an image. This value D is used as the other confidence indicator to reduce the confidence when this distance is small. The state of this confidence indicator is determined by:

$$C_i(l) = \begin{cases} \text{high confidence} & , \quad D > 80.0 \\ \text{medium confidence} & , \quad 50.0 < D \leq 80.0 \\ \text{low confidence} & , \quad D \leq 50.0 \end{cases} \tag{5.13}$$

5.3 Epipolar-Geometry Modality

This section describes the use of epipolar geometry for constraining the feature correspondence between two camera images. Given a set of corresponding feature points between two camera images, a point in one camera image defines an epipolar line in the other camera image on which the corresponding point lies. Epipolar geometry is an important concept when working with un-calibrated images and multiple viewpoints [59]. It has been widely used in structure and motion applications, such as scene reconstruction [167], motion recovery [147] and feature mapping for tracking [23]. In the following, firstly, an introduction to the theory of epipolar geometry is given in Section 5.3.1. Next, the application of epipolar geometry to the correspondence problem in matching subjects across cameras is presented in Section 5.3.2.

5.3.1 Epipolar Geometry

Perspective Epipolar Geometry

Figure 5.7 shows the epipolar geometry for two perspective cameras (ideal pinhole cameras). An introduction to the camera models is given in Appendix A. Figure 5.7 shows the two cameras indicated by their centres C_1 and C_2. The baseline connecting C_1 and C_2 intersects the image planes at the epipoles $\mathbf{e}$ and $\mathbf{e}'$. The 3D world point $\mathbf{P}$, C_1 and C_2 define an epipolar plane which intersects the image planes in the epipolar lines $\mathbf{v}$ and $\mathbf{v}'$. An image point $\mathbf{p}$ in image I_1 back-projects to a ray in the world defined by the image point $\mathbf{p}$ and camera centre C_1. This ray is imaged as the epipolar line $\mathbf{v}'$ in I_2 where the image point $\mathbf{p}'$, corresponding to the world point $\mathbf{P}$, must lie.

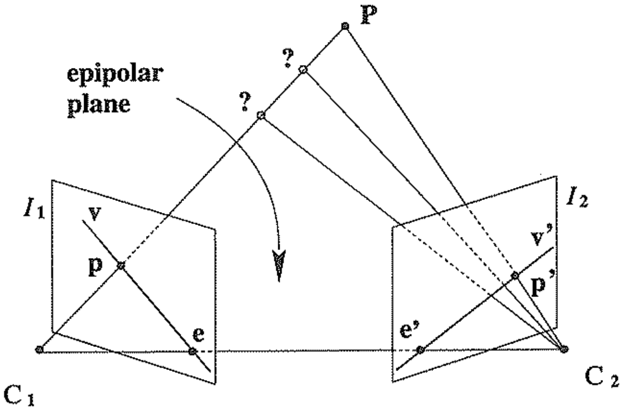


Figure 5.7: *Epipolar geometry for perspective cameras.*

For an arbitrary point $\mathbf{P}$ in the world, the image point $\mathbf{p}$, corresponding to $\mathbf{P}$, in the first image I_1 constrains the image position of $\mathbf{p}'$, corresponding to $\mathbf{P}$, in the second camera image I_2 on the epipolar line $\mathbf{v}'$. This line is namely the intersection of epipolar plane and I_2. Note that for epipolar geometry the image point $\mathbf{p}$ back-projects to a ray in the 3D world and there is no constraint on this point and the world point could possibly lie at any point along this ray. Whereas in the case of homography, the ray is constrained on a scene plane where the intersection of this ray and the scene plane is the world point $\mathbf{P}$.

Affine Epipolar Geometry

The affine camera model provides a good approximation of the perspective model when the FOV (Field Of View) is small and the variation in depth of the scene along the line of sight is small compared to its average distance from the camera [146]. Figure 5.8 shows the epipolar geometry of two affine cameras. The diagram on the left shows that all projection rays are parallel and perpendicular to the image plane since the optical centre of an affine camera lies at infinity. An image point **p** in image I_1 back-projects to a ray in the world defined by the camera centre C_1 and **p**. This ray is imaged as the epipolar line $\mathbf{v}'$ in I_2, so the image of world point **P** in I_2 must lie on $\mathbf{v}'$. The diagram on the right shows that all epipolar planes are parallel and hence so are the epipolar lines. This is because the optical centre of an affine camera lies at infinity, all projection rays are parallel. Thus, the affine camera preserves parallelism.

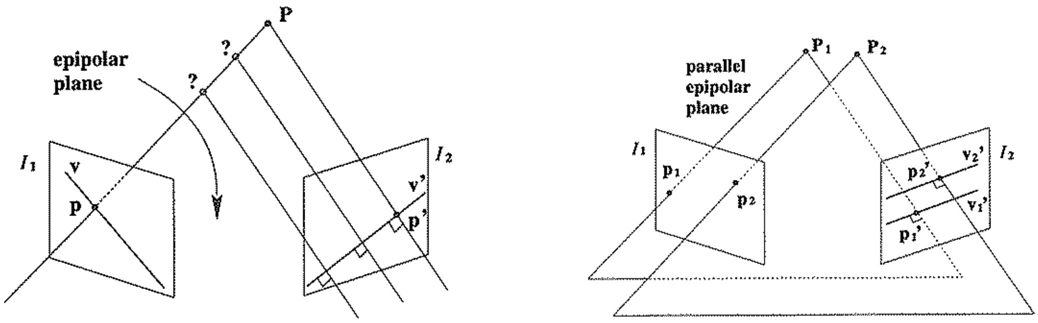


Figure 5.8: *Epipolar geometry for affine cameras.*

This work has used the affine camera model rather than the more familiar perspective model. The use of this model allows the system to use the distance between parallel epipolar lines as a modality confidence indicator (discussed in Section 5.3.2). Since the affine camera model only provides a good approximation of the perspective model when the FOV is small and the variation in depth of the scene along the line of sight is small compared to its average distance from the camera [146]. In the office where the experiment conducted, the modality may become less reliable when the depth variation is large compare to the average distance to the camera (see an example in Figure 5.11). The success or failure of applying this affine epipolar geometry modality for inter-camera subjects correspondences depends on the distance between the epipolar lines and the number of people in the scene. For example, when people (of about the same height and depth) are close to each other, the computed epipolar lines will also be close to each other. This can result in incorrect matches, if the affine camera assumption does not hold (e.g. people are close to the cameras).

Fundamental Matrix

The *fundamental matrix*, $\mathbf{F}$, needs to be estimated in order to compute the epipolar line across camera images. The fundamental matrix contains the geometric information of epipolar geometry between two views. It satisfies the condition that for any pair of corresponding points $\mathbf{p} \leftrightarrow \mathbf{p}'$ in the two images [59]:

$$\mathbf{p}'^{\mathrm{T}}\mathbf{F}\mathbf{p} = 0. \tag{5.14}$$

This is because if point $\mathbf{p}'$ corresponds to $\mathbf{p}$, then $\mathbf{p}'$ lies on the epipolar line $\mathbf{v}' = \mathbf{F}\mathbf{p}$. Since in a homogeneous coordinate system, the inner product of a point and a line, containing that point, is equal to null. In other words, $0 = \mathbf{p}'^{\mathrm{T}}\mathbf{v}' = \mathbf{p}'^{\mathrm{T}}\mathbf{F}\mathbf{p}$. The fundamental matrix depends only on the relative pose (i.e. position and orientation) between two cameras and does not depend on the scene structure [99]. Note that this fundamental matrix can be obtained from point correspondences alone. The fundamental matrix can also be computed based on camera calibration. In this case, the fundamental matrix reduces to the *essential matrix* [99]. For the affine camera model, Equation (5.14) can be expressed as [146, 59]:

$$\mathbf{p}'^{\mathrm{T}}\mathbf{F}_a\mathbf{p} = \begin{bmatrix} x'_i & y'_i & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & a \\ 0 & 0 & b \\ c & d & e \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = 0, \tag{5.15}$$

where the subscript, $_a$, in $\mathbf{F}_a$ is indicative of the use of affine camera model, $\mathbf{F}_a$ is the affine fundamental matrix, $(x'_i, y'_i, 1)$ and $(x_i, y_i, 1)$ are the corresponding points in homogeneous coordinates related to $\mathbf{p}_i = (x_i, y_i)$ and $\mathbf{p}'_i = (x'_i, y'_i)$ in the two images, a, b, c, d and e are the elements in $\mathbf{F}_a$. From Equation (5.15), each point match in two images (i.e. a point in I_1 and its corresponding point in I_2) gives rise to one linear equation:

$$ax'_i + by'_i + cx_i + dy_i + e = 0. \tag{5.16}$$

This equation, called the *affine epipolar constraint* [146], is defined up to a scale factor, so $\mathbf{F}_a$ can be computed uniquely from only 4 point correspondences, provided the 3D points are in general positions which do not violate the conditions where epipolar geometry is undefined (discussed below). To estimate $\mathbf{F}_a$, more than 4 points are used with the SVD method [124]. Some optimisation methods can be found in [146, 59] that minimise a cost function based on image distance. Given the fundamental matrix $\mathbf{F}_a$, the epipolar line in I_2, corresponding to image point $\mathbf{p} = (x_i, y_i)$ in I_1, can be represented as

$$\mathbf{v}' = \mathbf{F}_a\mathbf{p} = (a, b, cx_i + dy_i + e)^{\mathrm{T}}, \tag{5.17}$$

where the vector $(k_1, k_2, k_3)^{\mathrm{T}}$ is the representation of a line $k_1x + k_2y + k_3 = 0$ in a homogeneous coordinate system. Note that there are two conditions where epipolar geometry is undefined [167]:

1. when all scene points are coplanar, and
2. when two images are from a rotated camera or from two cameras with a common camera centre.

However, in both conditions, the image correspondences are defined by homography.

5.3.2 Using Epipolar Geometry for Inter-Camera Subject Correspondences

To obtain the epipolar geometry for our two-camera MCCT setup, a set of corresponding 3D points in two camera images are required to estimate the affine fundamental matrix $\mathbf{F}_a$. A set of corresponding points is obtained by extracting the feature points from multiple pairs of image sequences from two cameras. In each pair of image sequences, a single person is seen walking around the office, and the highest points of this person are extracted from two camera images. From multiple pairs of sequences with different people of different heights, a set of corresponding points is obtained for computing $\mathbf{F}_a$. To apply affine epipolar geometry for matching, the highest point of the subject is used as the feature point to represent the image position of a subject. Given the estimated fundamental matrix $\mathbf{F}_a$ (see Equation (5.15)), the highest point $\mathbf{p} = (x_i, y_i)$ of a subject in a camera image, I_1, is used to compute its associating epipolar line as $ax + by + (cx_i + dy_i + e) = 0$ (see Equation (5.17)) in the other camera image, I_2. The computed epipolar line is then used to search for the corresponding subject in I_2.

Figure 5.9 shows an example of the computed epipolar lines across camera images. The highest points, marked with '+', of two subjects in the right camera image are used to compute the epipolar lines $\mathbf{v}_{\mathrm{A}}$ and $\mathbf{v}_{\mathrm{B}}$ in the left camera image. The corresponding points should lie on these lines. Therefore, the distance between the computed epipolar line and highest point of the candidate matching subject is used as a match score. Based on a comparison of the distance between the highest points (marked with '$\circ$') of

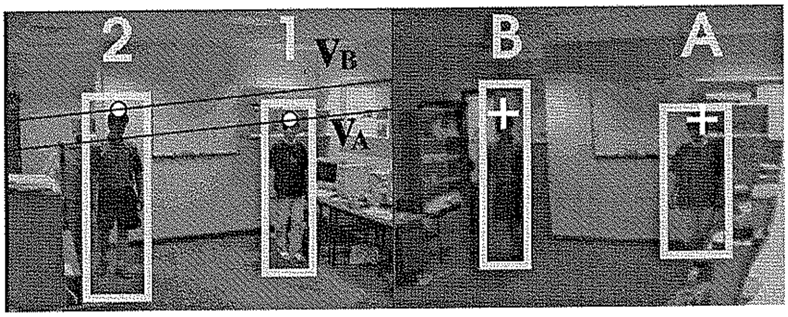


Figure 5.9: *An example of the use of epipolar geometry for matching subjects across camera images.*

the subjects and the epipolar lines, subject 1 (S_1) is matched to S_A and S_2 to S_B. In this example, the subjects in two camera images are correctly matched. However, the epipolar-geometry modality can be less reliable.

Figure 5.10 shows an example of incorrect matches caused by the ambiguous position of extracted feature point. Two people are visible in both camera images. The highest points, marked with '+', of two subjects in the right camera image are used to compute the epipolar lines $\mathbf{v}_{\mathrm{A}}$ and $\mathbf{v}_{\mathrm{B}}$ in the left camera image. The highest points, marked with '∘', of the corresponding subjects should lie on these lines. Based on a comparison of the distance between the highest points and the epipolar lines, the subject correspondences between two camera images can be determined. Due to an incorrect segmentation of the highest point of subject A, S_A, the related epipolar line $\mathbf{v}_{\mathrm{A}}$ is closer to the highest point of S_2. Consequently, the ambiguous position of the extracted feature point results in incorrect matches: S_A is matched to S_2 and S_B to S_1. In such a case, the BBN can adaptively reduce the modality confidence (see Section 5.3.2) in order to reduce the relative influence on the subject correspondences between two camera images. This adaptive adjustment of modality confidence is achieved by using the modality confidence indicator based on the segmentation status of the highest point. Note that the orientation of the epipolar line is determined by the relative motion between cameras [189].

Figure 5.11 shows an example of the case where the computed epipolar line is less reliable. The highest point, marked with '+', of the subject in the right camera image is used to compute the epipolar line in the right camera image. The computed epipolar line does not pass through the extracted highest point, marked with '∘', of the subject in the left camera image. This is because the affine camera model only provides an good approximation of the perspective model when the FOV is small and the variation in depth of the scene along the line of sight is small compared to its average distance from the camera [146]. In this case, the person just enter the overlapping FOVs of two cameras and are close to the right cameras. The depth of this person is smaller compares to the average distance from the right cameras such that the assumption of the camera model does not hold.

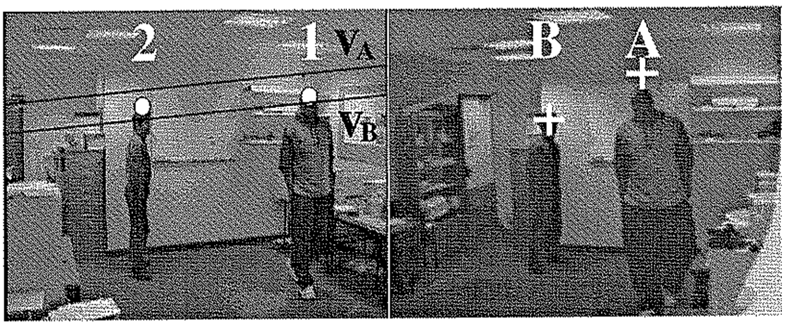


Figure 5.10: *An example of incorrect matches caused by the ambiguous position of an extracted feature point.*

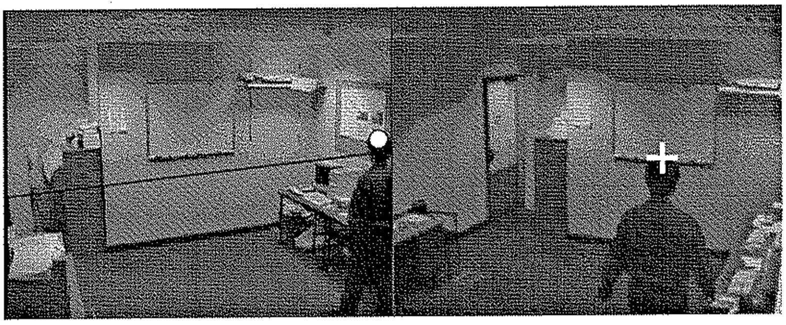


Figure 5.11: *An example of the case where the affine epipolar geometry modality is less reliable when the assumption of the affine camera model does not hold.*

Ideally, in one camera image, the highest point of the corresponding subject should lie on the epipolar line which was computed from the highest point of the subject in the other camera image. Therefore, the likelihood of a candidate match should be a decreasing function of one's related distance x between the highest point and the epipolar line. Such a distance is assumed, again, as a Gaussian variable with zero mean and defined as a probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{x^2}{2\sigma^2}). \tag{5.18}$$

The likelihood of a subject in I_2, with a distance x, being the corresponding subject in I_1 is determined by the value of the above density function. The comparison between subjects is based on the MD, $\mathcal{M}_e = \frac{x^2}{\sigma^2}$. Similar to the homography modality (see Equation (5.8)), the accumulated information:

$$M = \frac{1}{\sum_{i=0}^{q-1} \alpha_i} \sum_{i=0}^{q-1} \alpha_i \mathcal{M}_e(l-i), \tag{5.19}$$

is used in the comparison node of epipolar modality for determining the inter-camera subject correspondences, where q is the number of frames of accumulated evidences, l is the frame index and α_i is the weight used for setting more recent evidence with higher weights.

To determine the comparison result (i.e. similar or not similar) between each pair of subjects based on epipolar geometry, Equation (5.10) is used again. In this case, $\mathcal{X}_T^2$ is set to 3.84 corresponding to a 95% probability of finding the true feature for $n_f = 1$ (i.e. dimensionality of the feature vector).

The modality confidence indicator for the epipolar-geometry modality is defined as the mean distance, D (pixels), between each pair of affine epipolar lines in an image. These lines are computed from the highest points of all subjects in the other image. When the mean distance is shorter, the confidence is set lower. The state of this confidence indicator is determined by:

$$C_i(l) = \begin{cases} \text{high confidence} & , \quad D > 20.0 \\ \text{medium confidence} & , \quad 10.0 < D \leq 20.0 \\ \text{low confidence} & , \quad D \leq 10.0 \end{cases} \tag{5.20}$$

Moreover, the segmentation status of the highest point is also used to indicate the confidence (as defined in 5.2.2). The confidence is set lower when the positions of the highest points between consecutive frames change suddenly. This is because the image position of the highest point is used to determine the subject correspondences in the modality of epipolar geometry. When the position of the highest point is less reliable, the subject correspondences can also be less reliable.

5.4 Landmark Modality

The main reasons for using landmarks (easily recognisable scene structures) in computer vision are the low hardware costs and simple computation [184]. There are a lot of landmark-based position estimation techniques used in computer vision, e.g. using an angle measurement plus range data for determining the position of a robot. Landmarks are also widely used in autonomous robot navigation, e.g. [142, 143]. A good review of position estimation techniques based on landmarks can be found in [161].

Most of these methods focus on self-localisation for determining the position directly based on sensor measurements, such as camera, sonar, laser and infrared sensor. Since the correspondence of subjects between two camera images is defined by the 3D scene structure, this work aims to use prior knowledge of sparse scene landmarks relative to the cameras to constrain the correspondence. This is achieved by using the position of a subject relative to the landmarks in one camera image to geometrically reason about the position of corresponding subject in the other camera image. This constraint based on landmark modality is computationally a very efficient algorithm. However, the people tracked in the environment must be viewed in front of the landmarks and in the overlapping area of FOVs. Prior knowledge of the image positions of the line landmarks in two camera images is necessary, but only needs to be obtained once off-line. To build the correspondence of these lines between two camera images, the method proposed by Schmid and Zisserman [141] can be used which applies epipolar geometry to reduce the matching complexity. The correspondences of the line landmarks is built by hand in this work. The following description of how to determine inter-camera subject correspondences based on the landmark modality is a novel approach proposed in this thesis [28].

5.4.1 Multiple Camera Images and Landmarks

The goals of using the spatial reasoning method based on landmarks are twofold:

1. To determine the relative world position (cell, as discussed later) of a subject by using visual information from two camera images together.
   The idea behind this position estimation is that in general, even though the range data is not available from un-calibrated cameras (where the range data can be used to locate an object and obtain an absolute position, e.g. (x,y) coordinates in the world), the relative world position of a subject with respect to the landmarks

and static cameras can still be determined. This relative position is captured by finding the image positions of the object with respect to the landmarks in both camera images.

2. To use the correlation of the relative positions of subjects between two camera images as a constraint on the inter-camera subject correspondences.

In the following, the capture of the relative position with respect to the cameras and landmarks in a single camera image is described, before an explanation is given of how to use information from two camera images for achieving these two goals.

Landmarks in One Camera Image

Figure 5.12 illustrates multiple vertical line landmarks on the wall. A line landmark in the world is projected onto an image plane as a line. The projected image line is defined by a projection plane through the line landmark in the world and the camera centre. The space circumscribed by the four projection planes, which are defined by the two neighbouring line landmarks, the top and the bottom boundary of a camera image and its camera centre, is called a *Vertical Volume* (VV) of a camera. The area on an image plane, corresponding to a VV, is called a *Vertical Area* (VA) of a camera. For example, the VV, marked with bold lines on its boundary, corresponds to the 3rd VA in the right camera image. The number of the VA (shown as numbers on top of the camera images in Figures 5.12-5.14) is defined such that it increases from left to right and begins at 1. Through the imaging process, a subject in a VV is projected to the corresponding VA on the image. This process creates potential ambiguity due to the fact that one dimension is lost in the 3D to 2D projections of the spatial scene. As a consequence, for a person viewed in a VA, one can only infer the world position of this person in the whole corresponding VV, without knowing the absolute position.

Landmarks in Two Camera Images

Figure 5.13 shows two cameras with their overlapping FOV partitioned by the projection planes into different small subspaces. Each subspace is the intersection of two VVs of different cameras. These subspaces are called *cells*. By back-projecting a VA in both camera images, two corresponding VVs (see Figure 5.12) intersect in the 3D space and the intersection defines a cell in the scene. Due to the ambiguity mentioned above, Figure 5.14 shows that the person appearing in the second VA, l_2, of the left camera image may correspond to a subject in the first VA, r_1, or the second, r_2, in the

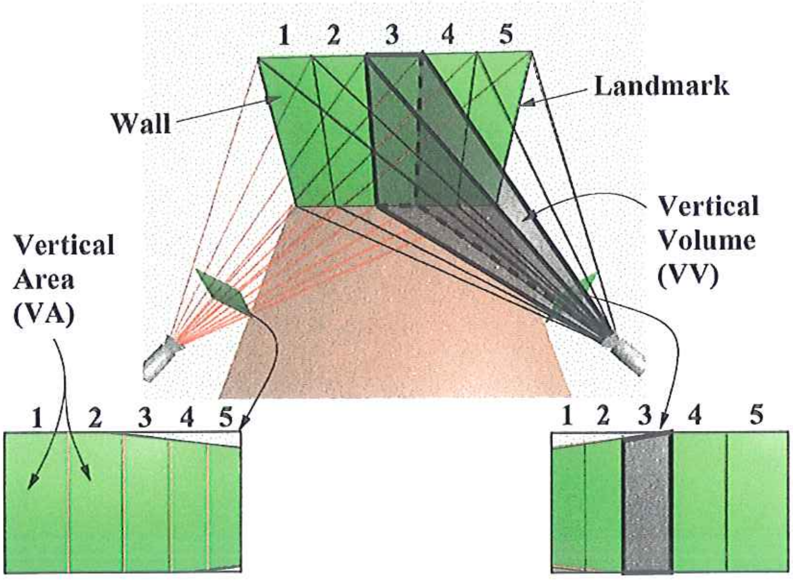


Figure 5.12: *The vertical line landmarks in the scene.*

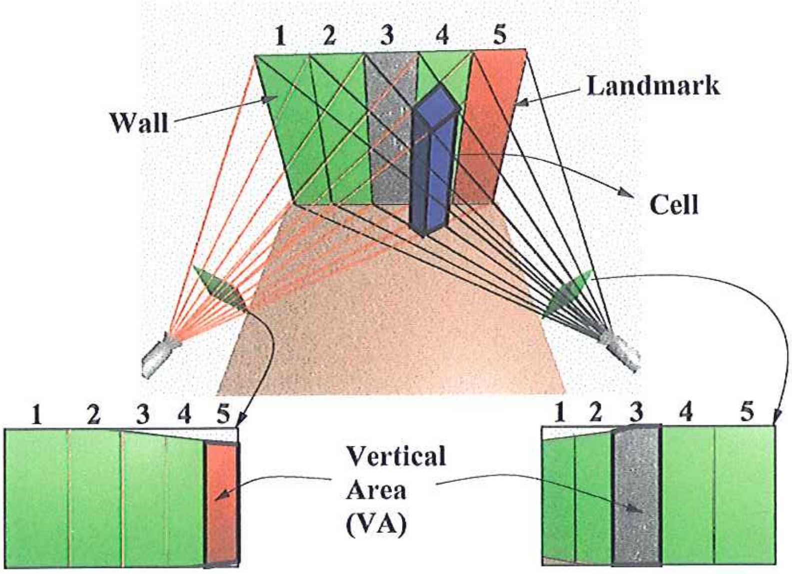


Figure 5.13: *The cells in the overlapping FOVs of two cameras.*

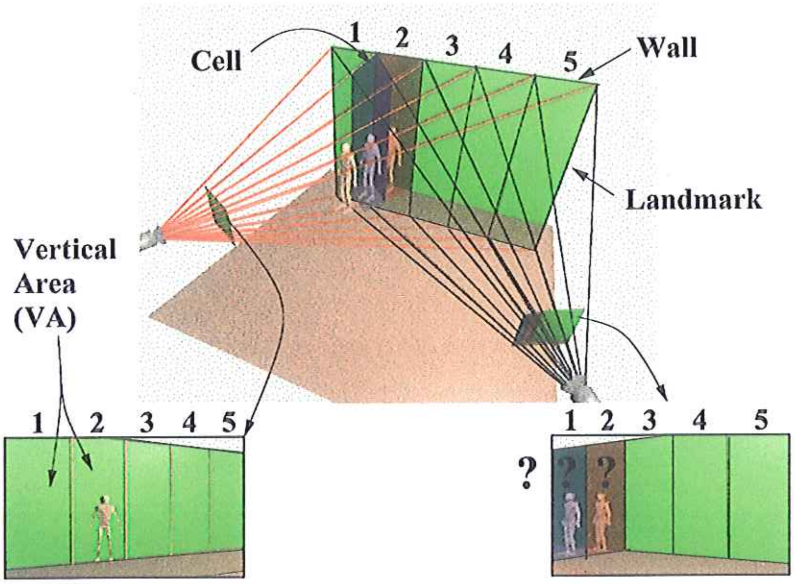


Figure 5.14: *An example of the ambiguity of the world position of a person.*

right camera image, or may not even appear. Based on this knowledge, we define the following two rules for reasoning about the image positions of corresponding subjects across cameras assuming the subjects are in the overlapping FOVs [28]:

Rule 1. For a subject S_1 imaged in the left camera image within the i^{th} VA, l_i, and a subject S_A in the right camera image with the j^{th} VA, r_j,

- if S_A corresponds to S_1, then $j \leq i$ and $V_a = r_1 \cup \cdots \cup r_i$.

V_a is the constrained area of the corresponding subject of S_1 in the right camera image and this area is called *valid area*. Note that the reverse of the rule is not necessarily true, so any subject in the valid area in the right camera image could correspond to S_1.

Rule 2. For a subject S_A imaged in the right camera image within the j^{th} VA, r_j, and a subject S_1 in the left camera image with the i^{th} VA, l_i,

- if S_1 corresponds to S_A, then $i \geq j$ and $V_a = l_j \cup \cdots \cup l_{max}$.

V_a is the valid area of the corresponding subject of S_A in the left camera image and the l_{max} is the right-most VA in the left camera image.

5.4.2 Using Landmarks for Inter-Camera Subject Correspondences

To apply landmarks for the purpose of matching subjects across cameras, the highest point is again used to represent the image positions of subjects. From the relative position of the highest points with respect to the landmarks, the subject correspondences between two camera images can be reasoned about by applying the rules defined above.

Figure 5.15 shows an example where the correspondence of two subjects between two camera images can be built correctly based on landmark modality. Two subjects with their highest points marked with '+' are seen in both camera images with different VAs (Vertical Areas) indicated on top of the images. By ***Rule 1***, subject B, S_B, with VA r_1 is matched to S_2 with VA l_2 since the valid area in the right camera image for matching S_2 is the union of VA r_1 and r_2. Both S_A and S_B can correspond to S_1 since both are in the valid area. Similarly, by ***Rule 2***, S_1 is matched to S_A, and both S_1 and S_2 could correspond to S_B. Consequently, by considering both rules S_A is matched to S_1, and S_B to S_2.

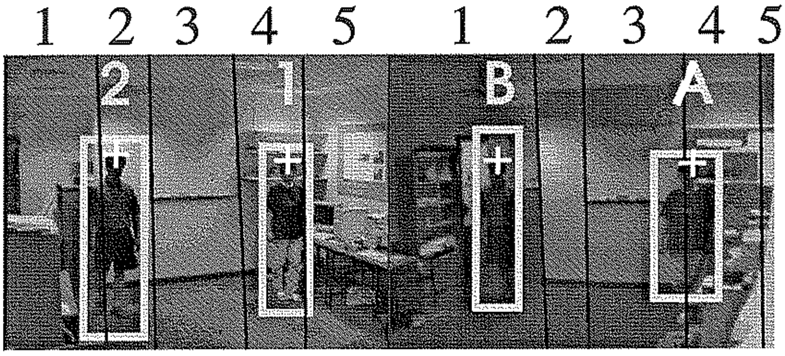


Figure 5.15: *An example of the use of line landmarks for matching subjects across camera images.*

Figure 5.16 shows an example where the landmark modality cannot be used to find a unique matching subject across camera images. Two subjects with their highest points marked with '+' are seen in both camera images and with different VAs (Vertical Areas) indicated on top of the images. By ***Rule 1***, the valid area in the right camera image for matching S_2 is the union of r_1 and r_2. Since both S_A and S_B are in this valid area, no unique subject can be matched to S_2. A similar situation arises in determining the corresponding subject in the right camera image for S_1. This situation also occurs when applying ***Rule 2***. Therefore, in this example, no match can be made based on the landmark modality. Such a matching ambiguity exists because there is more than one subject in the valid area. However, those subjects which are not in the valid area can be ruled out from being the corresponding subjects.

After explaining the use of the landmark modality for determining the correspondence

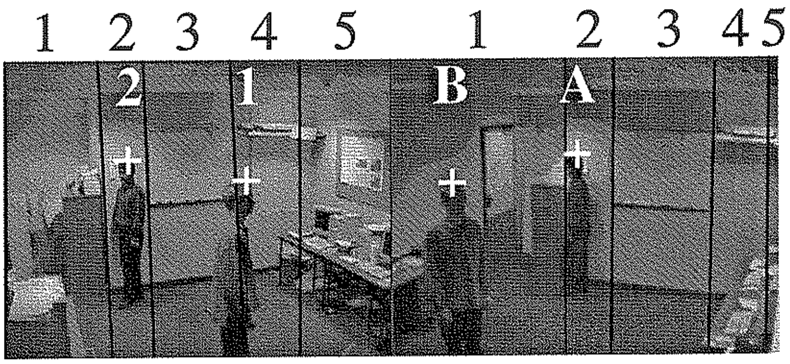


Figure 5.16: *An example of failure in using line landmarks for matching subjects across camera images.*

of subjects between two camera images, let us now consider the problem of data collection for landmark modality in the BBN. This landmark modality is modelled with two MUs (see Figure 4.10) to encode ***Rule 1*** and ***Rule 2*** respectively. Each MU compares all subjects in one camera image against all subjects in the other camera image. For example, each comparison node in the MU related to ***Rule 1*** compares a subject in the left camera image against all m subjects in the right camera image. Thus, this MU compares each subject in the left camera image against all subjects in the right camera image in m different comparison nodes based on ***Rule 1***. Since for a subject in one camera image, the subjects meeting the condition of the rule (i.e. within the valid area, V_a) in the other camera image, are all valid candidates, they should have the same probability of being the corresponding subject. Therefore, the matching probability of a subject with a VA number, $x = r_i$ (or l_i), is modelled with a uniform distribution:

$$f_m(x, l) = \begin{cases} 1/n & , \quad x \in V_a. \\ 0 & , \quad \text{otherwise.} \end{cases} \tag{5.21}$$

where l is the frame index and n is the number of subjects in the valid area of a camera image.

To compare subjects based on accumulated information, we define $g(l) = \frac{1}{n}$ for every frame and

$$G = \frac{1}{\sum_{i=0}^{q-1} \alpha_i} \sum_{i=0}^{q-1} \alpha_i g(l - i), \tag{5.22}$$

where q is the number of frames of the accumulated evidences, l is the frame index and α is the weight to set more recent evidence with higher weights.

The data collected in the comparison nodes and confidence indicator nodes for different modalities was explained. The collected visual evidence is fused by the BBN in order to probabilistically infer the correspondence of subjects between the two camera images. However, geometric modalities alone may not always provide enough constraints to match subjects across cameras. The next chapter will describe recognition-based modalities.

Chapter 6

Recognition-Based Modalities

6.1 Introduction

The previous chapter described the use of geometry-based modalities for matching subjects across camera images in order to perform MCCT. These modalities which used as local constraints on matching subjects across cameras are integrated over time by the BBN (Bayesian Belief Network, as described in Chapter 4). Since the use of geometry-based modalities alone may not provide enough constraints, a set of recognition-based modalities (i.e. apparent colour and apparent height) is also used. This chapter describes the recognition-based modalities including details of data collection for the evidential nodes of the BBN (i.e. the comparison node and confidence indicator node).

Since image variations between two widely-separated cameras can be highly significant (e.g. colour variations in two cameras, see an example in Figure 6.4), direct use of the extracted image features for building correspondence between two camera images does not always work. The novelty of the work presented in this chapter lies in the application of the Support Vector Regression (SVR) technique for learning the mapping of the visual information (i.e. apparent colour and apparent height) between two cameras [28]. The learnt mapping, encoding the correlation of subject appearance between two cameras is used to estimate subject appearance across cameras to compensate for the image variations for making inter-camera subject correspondences more reliable.

Figure 6.1 shows the block diagram of two phases for applying recognition-based modalities. The idea is to use the learnt mapping in the training phase to estimate the subject appearances for matching subjects across cameras during the tracking phase. The preprocessing in both training and tracking phases are the same. The preprocessing

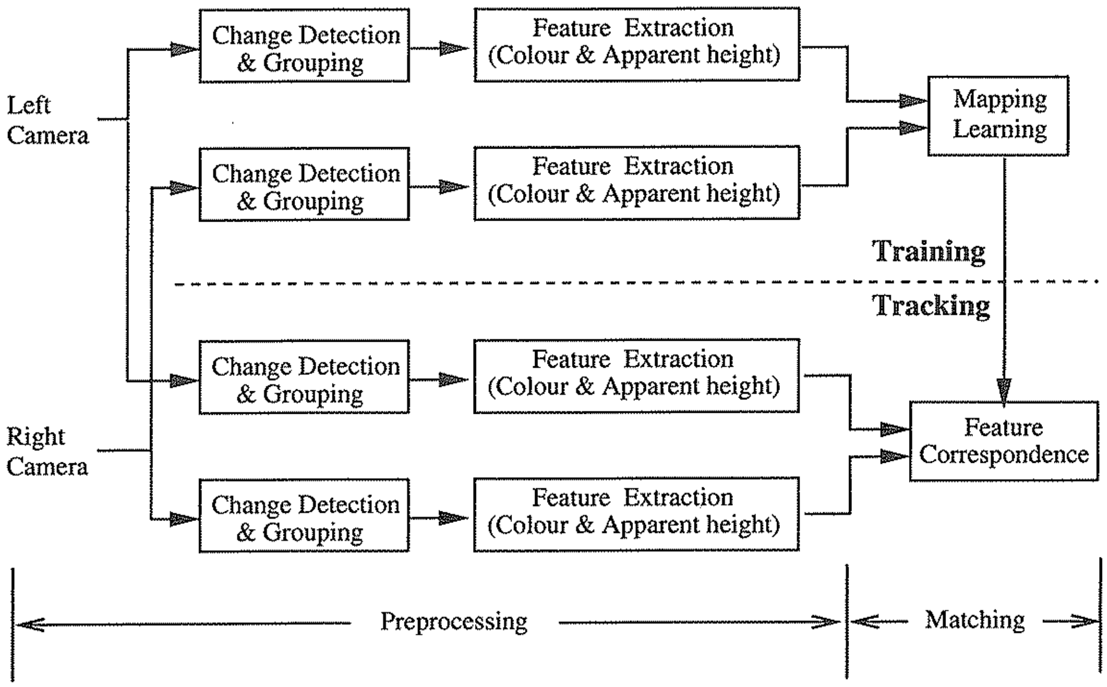


Figure 6.1: *Block diagram of matching the subjects' images across two camera images based on the learnt mappings of subject appearances.*

step includes two stages:

1. change detection and grouping, and
2. feature extraction.

The first stage was discussed in Section 3.2.1. The second stage will be discussed in this chapter.

This chapter is arranged as follows. First, the remainder of this section discusses general recognition problems and reviews some previous work related to the recognition-based modalities proposed in this thesis. Next, an explanation of feature extraction and appearance variations between two camera images is given in Section 6.2, and followed by Section 6.3 which gives a description of mapping learning and estimation of the subject appearances between two camera images, (the experimental results of mapping learning and appearance estimation will be given in Chapter 7). Finally, the data collection in BBN during the tracking phase for these recognition-based modalities is explained in Section 6.4.

Pattern Recognition

To match subjects between camera images, the image patterns extracted from the subjects in one camera can be used as a model to recognise (or search for) the corresponding subject in the other camera. Generally, recognition techniques focus on finding some distinctive patterns of interest and making decisions about the categories of the patterns. A good survey can be found in [73]. One of the main difficulties of image pattern recognition lies in the wide variation in the object's appearance due to changes in pose, scale and lighting condition such that the object appears differently from different viewpoints [159, 73].

To recognise objects in the images, different methods can be used, such as template matching [11], structural matching [32], statistical classification [73] and extraction of invariant features [5, 52]. Template matching uses a template (i.e. points, curves or shapes) to search for objects based on the similarity between two entities, while the structural matching method adopts a hierarchical perspective where a pattern is composed of simple sub-patterns. Statistical methods establish decision boundaries in the feature space to separate patterns belonging to different classes. Thus, an object class can be decided based on the probability distributions of the patterns belonging to the class. Invariant-feature methods attempt to extract object descriptions that remain constant under different geometric transformations or different illuminations [109], but may be only applicable to limited classes of objects, such as industrial parts [47]. Overall, the general problem of recognising complex patterns with arbitrary orientation, location and scale remains unsolved [73].

Previous Work

This thesis uses apparent colour and apparent height to match subjects in two camera images. Based on the learnt mapping, the extracted appearances of a subject in one camera image are used to estimate the appearances of the corresponding subject in the other camera image. In the literature, different features have been proposed to match people between images. For example, the apparent height ratio between consecutive frames is used to match subjects between consecutive images from a single camera [23]. The apparent colour is used for matching subjects in a single camera image [181] or multiple camera images [116, 158]. However, these different types of visual information cannot reliably be used for directly matching between multiple camera images without considering image variations. This is due to the colour shift between images, called the *colour constancy* problem, which can cause colour-based recognition to become less reliable [62]. On the other hand, the apparent height ratio between consecutive frames generally may not hold between multiple camera images due to the complex correlation of the apparent height between two camera images (e.g. lower part of a person is not visible in one image but visible in the other).

To match subjects in two camera images based on colour, one possible method is called *colour indexing* [160]. This colour-based recognition method computes the similarity between an image and a model colour histogram. As reviewed in Section 2.2.2, Stillman et al. [158] adopted this method to match subjects between the images from multiple cameras. The colour extracted from the subject image of one camera is directly used as a model to search for the corresponding subject in the image of the other camera. However, since illumination changes alter the observed image colours, colour indexing may not perform well enough under varying illuminations because it has no mechanism to handle colour shift [62]. Therefore, colour indexing may not be applicable for the purpose of matching subjects across cameras due to the significant colour shift between two camera images (see Figure 6.4).

To deal with this colour constancy problem, the *colour constancy methods* can be used. Since the image appearance can be affected by the reflectance of the surface, these methods aim to match objects colours under varying illumination by assigning some illumination-invariant descriptors to each object surface [20] (e.g. colour calibration method [29] and gamut mapping method [10]). Some good reviews of these algorithms can be found in [62, 20]. However, most of these colour constancy methods perform only on highly restricted images [20] and require some undesirable assumptions [62]. For example, the colour calibration method needs a standardised colour device (i.e. colour chart), and the gamut mapping method makes assumptions about image gamuts thus limits the possible number of illuminants in the scene [20].

In order to compensate for the colour shift for object recognition, an alternative approach is the sample-based method which learns the colour variation from a set of training samples. For example, Buluswar and Draper [20] learnt the colour variations of objects in outdoor scenes to make colour recognition more robust. This thesis applies the sample-based method to learn the colour mapping between two camera images. This mapping is then used to estimate the apparent colour of corresponding subjects across two camera images in order to compensate for the colour shift.

The other image feature used in this thesis for matching subjects across camera images is the apparent height. An alternative feature related to the height of human body is the true height of the subject [36]. The true heights can be extracted from each camera image used for inter-camera subject correspondences. However, this method, based on projective geometry, requires some assumptions to be made about the scene (i.e. parallel lines in the scene). Furthermore, the lens distortion can degrade the accuracy. The other method is to obtain the estimated true height from a camera image [115]. This technique compares the measured apparent height with a height histogram with respect to the image position of the subject. However, this method, similar to the previous method, cannot handle the situation where some body parts are not visible in the camera image.

The Approach: Sample-Based Method

To recognise an object in an image, one must have an internal model of how that object may appear [174]. There are a number of often difficult factors which can affect object appearances that must be considered in object recognition, e.g. camera parameters (including lens distortion and colour sensing characteristics), viewing geometry and illumination conditions. However, the underlying problem of matching subjects across cameras based on the image patterns is different from that of the recognition problem. The recognition problem needs a model to search for a certain object in an image where the imaging parameters are generally unknown. As such, the difficult factors mentioned above need to be considered explicitly in order to handle the image variations between the model and the object image. However, the task of matching subjects across cameras only involves finding the corresponding subjects in two camera images. In other words, if one can compensate for the subject appearance variation between two camera images, the subject correspondences can be established more reliably. The tracking task in this work is performed in an office environment where the factors are unknown but not entirely arbitrary (e.g. the cameras and the lighting conditions are fixed). As a consequence, the correlation of the subject appearances is also fairly stable. To compensate for the variations in subject appearances between two camera images, this thesis proposes to learn the correlation of the subject appearance between two camera images, and uses the learnt mapping to estimate the subjects' appearances across cameras. Thus, the difficult factors are implicitly handled in the learning process. Since the subject appearances generally vary with the world positions in the scene (see Figure 6.6 and

Figure 6.7), this correlation is learnt with respect to different sub-spaces (i.e. cells, see Figure 5.13) in the scene. By using the learnt mapping, one can estimate the visual appearances of different subjects across two camera images in order to obtain a suitable model for matching subjects. In this thesis, the mapping estimation of the subject appearances between two camera images is formulated as a regression problem (discussed in Section 6.3).

6.2 Appearance Variation between Two Camera Images

First, in Section 6.2.1, the feature extraction and representation is discussed, before a description of the representations of subject appearances used in this thesis is provided. Then, illustrations of the variations in appearance between the different images from the two cameras are given in Section 6.2.2.

6.2.1 Appearance Representation

The general effectiveness of an appearance representation lies in its robustness of identifying subjects in the images. The appearance of a subject's image depends on many imaging factors such as lighting conditions, viewpoint, articulation and geometric deformations of the object, and whether it is partially occluded by other objects [47]. It is therefore necessary to design subject representations which are robust to all image variations caused by these factors. Such representations can then be used for subject recognition in different camera images. Because colour, as mentioned in Section 3.2.1, is robust to common geometric distortions (such as rotation, translation, cropping, scaling) [14], it can be used as a reliable representation of subject appearance. Moreover, colour-based tracking complements spatial tracking and can be used for matching subjects in multiple cameras [116], though it cannot reliably handle the case when subject images are of a similar colour.

On the other hand, geometric descriptions can be used for identifying subjects and are generally robust with respect to illumination variations. However, most geometric feature-based recognition methods can only handle simple, flat, and rigid man-made objects [52]. For example, shape features are rarely adequate for discriminative recognition of 3D objects from arbitrary viewpoints in complex scenes [52], because natural objects viewed under realistic conditions do not have uniform shapes [159]. A simple geometric feature, apparent height [27], is used in this thesis to represent the subjects in the image for matching subjects across cameras. This is because a subject's apparent height is in general not affected by a person's orientation and holds a one-to-one relation between two camera images for a certain world position where the person is located.

Applying recognition-based modalities involves two stages of preprocessing (see Figure 6.1). The second stage is the feature extraction of the subject appearance (i.e. apparent colour and apparent height). In the following, the extraction and representation of subject appearance feature to be used for matching subjects across cameras are described.

Feature Extraction and Representation of Apparent Colour

The left image in Figure 6.2 shows the apparent colour extracted from the subject image. Certain domain knowledge (i.e. clothing is at a certain distance below the highest point of a person) is used to extract the sub-image from the segmented blob. This extracted sub-image may not correspond to the clothing of the subjects due to the ambiguous nature of the feature position (e.g. incorrect segmentation of the highest point). Some model-based methods can be used to analyse human motion and determine the pose of the subject in order to reliably extract the colour data from the blob (such as the carboard model used in the W^4 system [56] and the star skeleton model in [34]). To handle this problem, the confidence indicators in the BBN (see section 4.4.1) are used to adaptively reduce modality confidence in order to reduce the relative influence on the subject correspondences when the segmentation is less reliable (discussed further in Section 6.4.1).

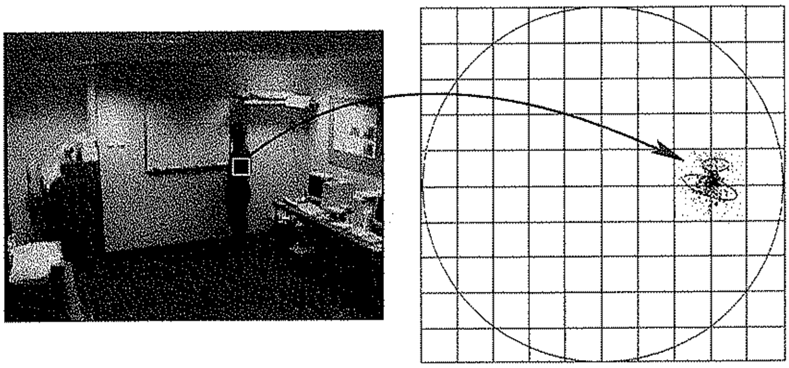


Figure 6.2: *An example of the extracted colour data from the subject's image and its Gaussian mixture model in HS-space. (Note that for convenience the polar coordinates of hue and saturation are drawn in Cartesian coordinates)*

To represent the colour, hue and saturation (HS) space is used for obtaining a limited level of intensity and robustness towards illumination changes by dropping the intensity component (i.e. value (V)). An introduction to Hue, Saturation and Value (HSV) colour space is given in Appendix E. The Gaussian mixture components used to model the colour distributions [185] are shown as elliptical contours of equal probability in HS space in the right graph of Figure 6.2.

Feature Extraction and Representation of Apparent Height

Figure 6.3 shows an example of the apparent heights of subjects in the two camera images. The top row shows two people viewed in both camera images. The bottom row shows the related binary foreground images of the segmented blobs and the apparent heights of the two subjects. The apparent height of a subject is defined as the image distance between the highest and the lowest points along the vertical direction of a segmented blob. It can be seen that the apparent height can be affected by the image boundary (e.g. subject 2 in the left camera image) and the objects in the scene due to body parts not being visible (e.g. subject 1 in the right camera image).

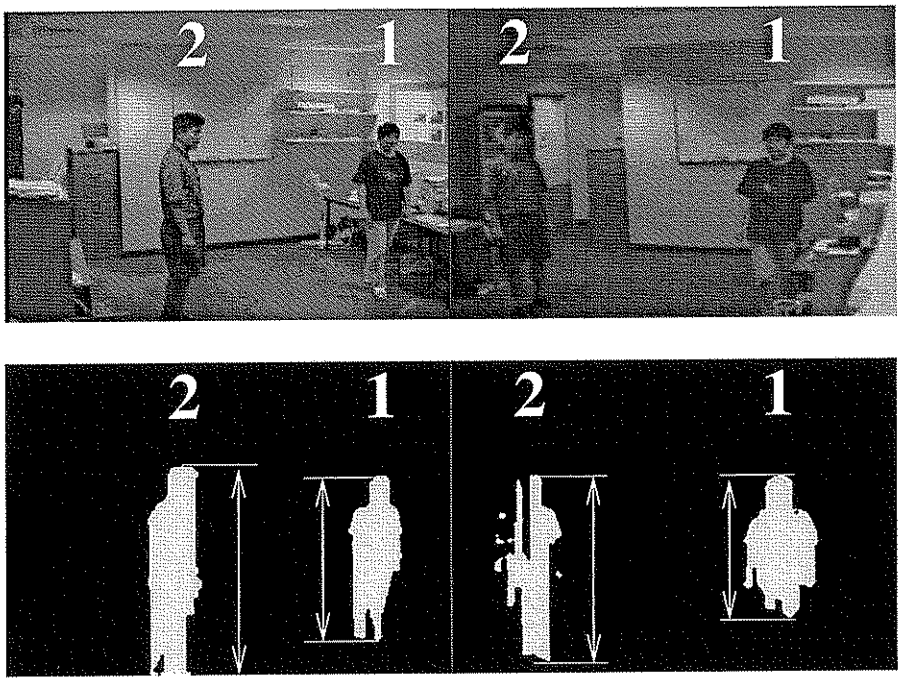


Figure 6.3: *An example of the apparent heights of subjects.*

Joint Features for Subject Representation

To represent subjects in our two-camera tracking system, a combination of different visual modalities from two cameras is used. We call this combined representation a *joint feature* [28]. A subject is represented with both spatial and appearance components (i.e. apparent colour and apparent height). This is because subject appearances are strongly correlated to their world positions (as illustrated in Section 6.2.2), the world position is incorporated in the subject representation for robust scene interpretation and understanding. Since the apparent colour and apparent height are generally independent, these two features are not further combined. Thus, both *joint colour feature* and *joint height feature* are used to represent a subject. Similar representations (incorporating spatial information) can be found in [122] where they are used to extract a meaningful description from satellite imagery. The author applied this joint representation further to represent subject images [181]. However, the spatial information which is used in their representation is the image position and the information used is from a single camera. Whereas in this thesis, the representation consists of the world position and uses the information from both cameras. The joint features are defined as:

- *joint colour feature*:

$$\mathbf{V}_c = (x_1, x_2, \mathbf{G}_1, \mathbf{G}_2), \tag{6.1}$$

- *joint height feature*:

$$\mathbf{V}_h = (x_1, x_2, h_1, h_2), \tag{6.2}$$

where x_i is the VA position in the image I_i, (x_1, x_2) is the cell position (see Figure 5.13), $\mathbf{G}_i$ is the 2D Gaussian variable used to model the apparent colour of subject image, and $\mathbf{G} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the mean on the HS plane $\boldsymbol{\mu} = (\mu_h, \mu_s)$ and a covariance $\boldsymbol{\Sigma}$.

6.2.2 Appearance Variation

Here, for both apparent colour and apparent height, an illustration of the variations in subject appearances between two camera images is given. This variation can make any direct use of apparent model for the purpose of matching subjects across cameras less reliable. These appearance variations are shown to be highly correlated to the world positions of subjects. As a consequence, one has to incorporate the world position for faithful estimation of the subject appearance across camera images.

Apparent Colour

To illustrate the colour variation between two camera images, a colour (red) sampled from the person's clothes, as shown previously in Figure 6.2, is shown in Figure 6.4 for the images from both cameras. The colour samples are from 400 frames of a sequence pair from two cameras where the person walks around an office. In each frame, only the mean of the colour samples in HS (Hue Saturation) space is plotted (see introduction to HS space in Appendix E). Each of the two clusters corresponds to the means of the colour distributions of 400 frames in a camera image. It can be seen that the colour shift in each camera image is significant. More significantly is the colour shift between the two camera images, with the two clusters being well separated. Therefore, direct use of the colour model obtained from one camera to search for the corresponding subject in the other camera image may be less reliable. This colour constancy problem in a multiple camera system must be considered in order to make the inter-camera subject correspondences more reliable.

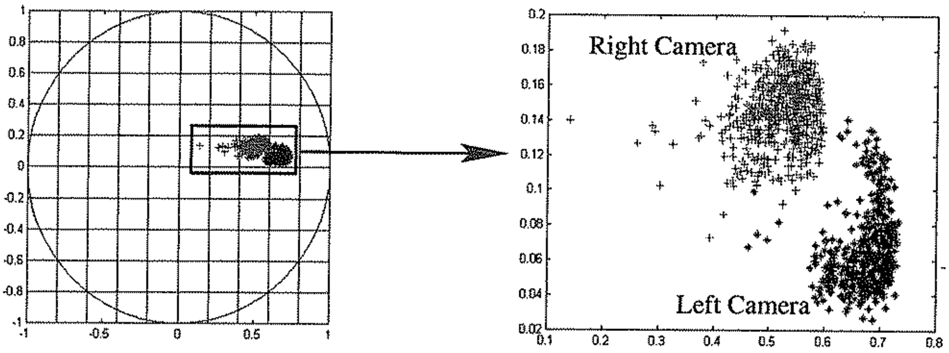


Figure 6.4: *An example of variations in apparent colours sampled from the images of a person in two cameras over 400 frames.*

Figure 6.5 illustrates a subset of the means of the colour distributions shown in Figure 6.4. This subset corresponds to the colour samples in two camera images when the person is in a given cell in the scene. The top row shows the means of the colour distributions in HS space in both images. The central graph in the bottom row illustrates the cell position (top view) where the person is located. The two side images in the bottom row show the related two VA positions in both images. These two VA positions define the cell position in the scene (see Figure 5.13). Compared to the colour samples corresponding to all the cells (as shown in Figure 6.4), it can be seen that the two clusters

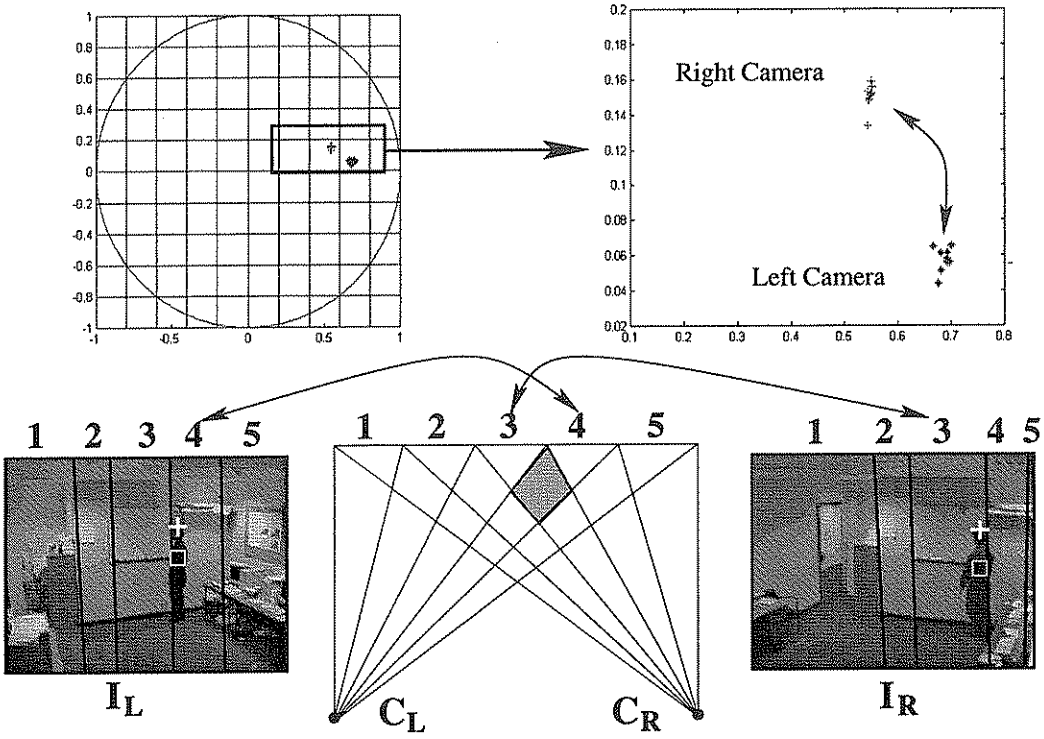


Figure 6.5: *An example of apparent colours sampled from a subject in two camera images over 12 frames when the subject is in the shaded cell's position. (Note that for convenience the polar coordinates of hue and saturation are drawn in Cartesian coordinates)*

(corresponding to the means of the colours in the two camera images when the person is in a certain cell) are much more compact. This indicates that the colour is highly correlated to the world position in the scene. Therefore, incorporation of the world position for learning the mapping of apparent colour in two camera images can make the mapping more reliable. Note that the areas of the cells in the central illustrative graph of the bottom row are not proportional to the actual areas in the 3D world. For example, the cell position defined by the 5^{th} VA in the left camera image and the 1^{st} in the right is much larger than other cells in the 3D world. This is because the 1^{st} VA in the right is much larger than other VAs. However, this cell position is shown in the central graph in the bottom row to be the smallest. Thus, the chances of the person in different cells during the sequence might not necessarily correspond to the cell areas in the central graph in the bottom row. Moreover, the person walks around the office randomly.

Figure 6.6 illustrates a subset of the means of the colour distributions of Figure 6.4. This subset corresponds to colour samples taken from images in the left camera image when the person is in different cells of the scene. The top row shows the means of the colour distributions in HS space of different frames. The bottom row illustrates the cell positions where the person is located. The arrows indicate the correspondence between the colour samples and the cells. The upper left graph shows the colour samples for three different cells and the upper right for four cells (by adding one to the three cells in the lower left graph). It can be seen that in the upper left graph, corresponding to the person positioned in three different cells, the colours for different cells are quite separated. This colour shift is caused by multiple illuminants in the scene and difference in viewing geometry (orientation of the subject surface normal with respect to the camera and the distance between them) and illumination geometry (orientation of the subject surface normal with respect to the illuminant and the distance between them) for these three cells. Now let us see the effect on the colour distributions if the cells (3D world positions where the person is located) are not as separated as the three in the lower left graph. The upper right graph corresponds to the addition of colour samples where the person is in the fourth cell. In such a case, the means of the colour distributions, corresponding to the fourth cell, can overlap with the means corresponding to the previous cells. This is because the fourth cell neighbours two of the three cells (as shown in the lower right graph). This indicates that if the cells are more separated, the colour distribution corresponding to different cells will be more separated.

Apparent Height

Figure 6.7 illustrates an example of variations in the apparent heights of a subject in two camera images. The upper left graph shows the apparent heights extracted from the subject in two images of two sequences captured from two cameras. The disparity between the apparent heights in two images is also shown. The apparent heights are extracted from the subject and shown over 380 frames from a sequence pair from two cameras where a person walks around the office. The apparent height drops significantly (at the 290^{th} frame) when the subject is in the cell position (shown in the upper right graph) defined by the two VAs in two images (see Figure 5.13) as shown in the bottom row. The apparent heights drop significantly in the right camera image due to the lower part of the subject not being visible.

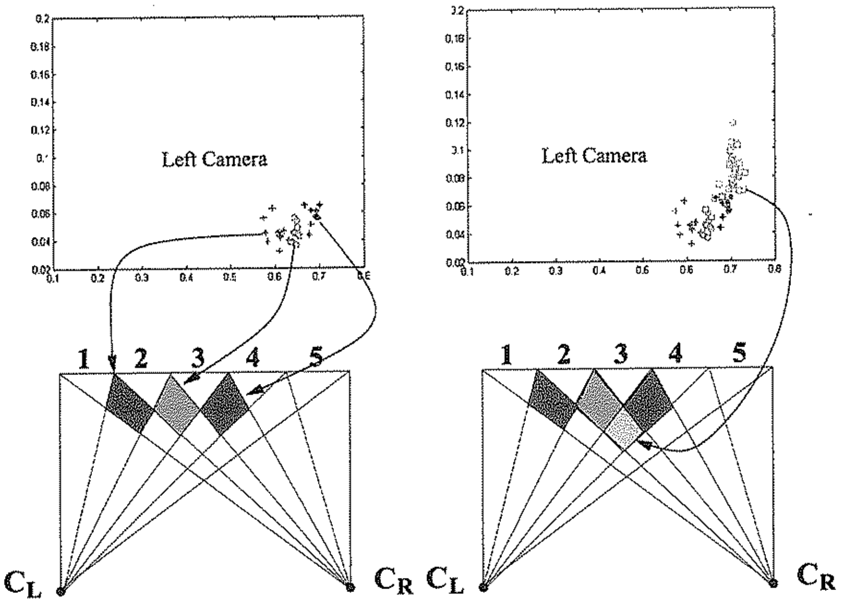


Figure 6.6: *An example of apparent colours sampled from the left camera image when the person is in different cells. (Note that for convenience the polar coordinates of hue and saturation are drawn in Cartesian coordinates)*

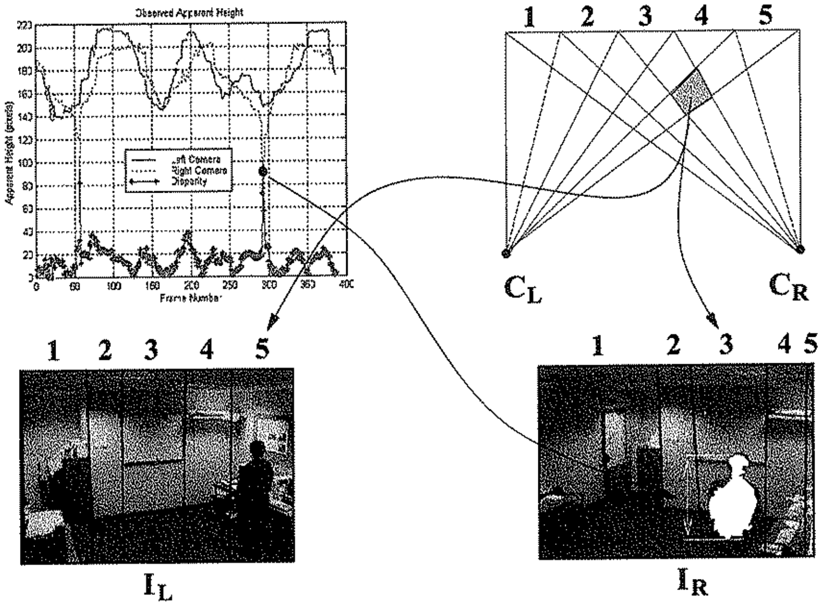


Figure 6.7: *An example of the variations in apparent height of a subject.*

The apparent height of a subject is not related to a person's orientation and generally depends on a person's height, viewing geometry (orientation of the object surface normal with respect to the camera and the distance between them), and camera parameters (e.g. lens distortion). One of the attractive characteristics of apparent height is that even with people of the same real height, their apparent heights in each camera image are not necessarily identical. This is because of perspective distortion and different depths of subjects. Figure 6.8 shows an example of the variations in apparent heights of subjects between two camera images. Two subjects are visible in both camera images. The real heights of both people are similar (about 178 cm). In the left camera image, the apparent heights of the two subjects are almost the same since the depth of these two subjects are almost identical. However, the apparent heights of the two corresponding subjects in the right camera image are not the same, because the depths of these two people are different with respect to the right camera. Thus, apparent heights can still be used for matching subjects across cameras even with people of the same real height. This phenomenon, and the one previously discussed (as shown Figure 6.7), indicate that the apparent height taken from both camera images together can be a strong cue to the world position of the subject. In fact, the apparent heights in two camera images are highly correlated to the world position of the subject.

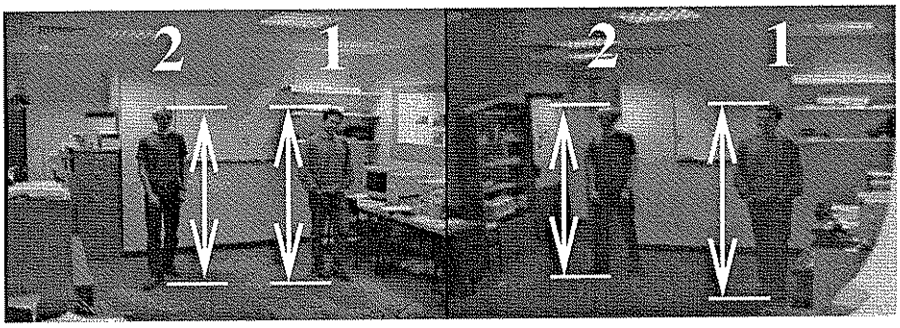


Figure 6.8: *An example of the variations in apparent heights of subjects between two camera images.*

6.3 Estimating Appearance Across Camera Images

After introducing the appearance-based representation and their variations, this section describes the training phase for learning the mapping of subject appearances between two cameras (see Figure 6.1). The learnt mapping is used to estimate subject appearances across camera images for matching subjects across cameras during tracking. Firstly,the selection of a set of training examples for different recognition-based modalities is described in Section 6.3.1. Next, the application of SVR (Support Vector Regression) for estimation of the mapping is performed in Section 6.3.2.

6.3.1 Training Joint Appearance Features

Since the cameras in our system are static, the illumination conditions, illumination geometry and viewing geometry are fairly constant. Therefore, the correlations of both the apparent colour and apparent height of subjects between two cameras are also relatively constant. This thesis attempts to learn the mapping for these two modalities to capture the correlation of the appearance between two images [28]. The joint features including the appearances of a person in both camera images are used for representing a person. By incorporating the world position (i.e. cell) in the joint feature, the learnt mapping can encode the appearance variances between two images with respect to 3D positions.

To estimate the mapping of a person's appearance between two camera images, the joint features are extracted from the subjects' images of a sequence pair from two cameras where a person is walking around an office. To obtain the training set of joint colour features $\{\mathbf{V}_c\}$, each person wears clothes of a single colour and different persons wear different colours. A single Gaussian variable $\mathbf{G}_i$ is used to model the apparent colour of the person's clothes in each image I_i of the two cameras to obtain $(\mathbf{G}_1, \mathbf{G}_2)$. The 3D position (i.e. cell position, (x_1, x_2)) is obtained from extracting the VA x_i in each of the images I_i from the two cameras. From these two components, the joint colour features, $\mathbf{V}_c = (x_1, x_2, \mathbf{G}_1, \mathbf{G}_2)$, are found. From different sequences of a person wearing different colour clothes, the training set $\{\mathbf{V}_c\}$ is obtained. Similarly, a training set of $\{\mathbf{V}_h = (x_1, x_2, h_1, h_2)\}$ from different sequences of different people with different real heights is obtained.

6.3.2 Mapping Using Support Vector Regression

In order to estimate the nonlinear mapping of the appearance of a subject between two camera images, the mapping estimation is formulated as a regression problem and the SVR (Support Vector Regression) method [44] is used. A brief introduction to SVR is given in Appendix F. Unlike some other regression techniques, SVR has some desirable properties including:

- It is not necessary to determine the model structure before training [154]. The final regression function can be expressed by using a set of "important examples" called *Support Vectors* (SVs).
- The regression estimation problem can be solved as a quadratic optimisation problem. It is guaranteed to converge to the global optimum of the given training set.
- By introducing a kernel function, the nonlinear regression function is implicitly defined by a linear combination of training examples (i.e. SVs) in a high-dimensional feature space.

The task of learning a nonlinear mapping function is described as follows. Given a set of training examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ with input patterns $\mathbf{x}_i \in \mathcal{R}^n$ and interpretation $y_i \in \mathcal{R}$, our goal is to find the function, $f(\mathbf{x})$, that has the most ε tolerance from the actual interpretation, y_i, for all the training data. For the mapping of apparent colours ($\mathbf{G}_1$ and $\mathbf{G}_2$) of the same entity in two camera images, I_1 and I_2, our experiments show that the estimated covariance $\mathbf{\Sigma}$ is less reliable. Therefore, only the mean position, $\boldsymbol{\mu} = (\mu_h, \mu_s)$, of the apparent colour of the corresponding subject in one camera image is estimated. This is based on the observed colour model in the other camera image. For example, for the mapping of apparent colour from image I_1 to I_2, two mappings are learnt by setting the input patterns to $\mathbf{x}_i = (x_1, x_2, \mathbf{G}_1)$, and the interpretations to $y_i = \mu_h$ and $y_j = \mu_s$ respectively. Similarly, for mapping the apparent height from I_1 to I_2, a mapping is learnt by setting $\mathbf{x}_i = (x_1, x_2, h_1)$ and $y_i = h_2$.

The SVR problem can be formulated as a quadratic programming problem by maximising [44, 154]:

$$\begin{aligned} W(\alpha^*, \alpha) &= -\frac{1}{2}\textstyle\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(\mathbf{x_i}, \mathbf{x_j}) \\ &\quad -\varepsilon\textstyle\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i), \qquad (6.3) \\ \text{subject to} &\quad \textstyle\sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0, \qquad (6.4) \\ &\quad 0 \le \alpha_i^*, \alpha_i \le C, \qquad (6.5) \end{aligned}$$

which provides the solution

$$f(\mathbf{x}) = \textstyle\sum_{i=1}^{l}(\alpha_i^* - \alpha_i)K(\mathbf{x}, \mathbf{x_i}) + b, \qquad (6.6)$$

where α_i and α_i^* represent the parameters of the learning machine, K is the kernel function and C is a penalty factor and b is a threshold.

Two SVR-based apparent colour mappings (one for μ_h and the other for μ_s) are constructed to estimate the apparent colour in one camera image from the other camera image, and another one mapping constructed for apparent height (h). These mapping are learnt from different apparent colours and different people with different real heights (see Section 7.2.1). One drawback of this SVR method is that it may be computationally very expensive [153].

6.4 Using Estimated Appearance for Inter-Camera Subject Correspondences

Having described how to learn the mapping of subject appearances between two camera images, the use of the estimated appearance for constraining subject correspondences between two camera images is explained in this section.

6.4.1 Using Estimated Apparent Colour for Correspondences

To match subjects in two camera images based on apparent colour, the learnt mapping is used to estimate the apparent colour of the corresponding subjects across cameras. For a subject with a VA (Vertical Area, see Figure 5.12) x_1 and apparent colour $\mathbf{G}_1$ in image I_1, the goal is to search for the corresponding subject in image I_2. In doing so, the learnt mapping is used to estimate $\boldsymbol{\mu}'_2 = (\mu'_{h2}, \mu'_{s2})$ for each subject in I_2 (with a VA, x_2, and observed apparent colour $\mathbf{G}_2 = (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$), based on the observations $(x_1, x_2, \mathbf{G}_1)$. Combination of estimated mean, $\boldsymbol{\mu}'_2$, and covariance, $\boldsymbol{\Sigma}_2$, (i.e. $(\boldsymbol{\mu}'_2, \boldsymbol{\Sigma}_2)$) is used as a colour model to compute the likelihood of a candidate matching subject in I_2. Thus, the conditional probability of a pixel λ in a subject image in I_2 being the subject, $\mathcal{S}$ in I_1 (modelled as a mixture with u components), is given as:

$$p(\lambda|\mathcal{S}) = \sum_{i=1}^{u} p(\lambda|i)P(i), \tag{6.7}$$

where $P(i)$ is the prior probability that the pixel λ was generated by the i^{th} component, $\sum_{i=1}^{u} P(i) = 1$. Each component is a Gaussian with mean $\boldsymbol{\mu}'_2$ and covariance matrix $\boldsymbol{\Sigma}_2$, and:

$$p(\lambda|i) = \frac{1}{2\pi|\boldsymbol{\Sigma}_2|^{1/2}} exp(-\frac{1}{2}(\lambda - \boldsymbol{\mu}'_2)^T \boldsymbol{\Sigma}_2^{-1}(\lambda - \boldsymbol{\mu}'_2)). \tag{6.8}$$

To compare the candidate matching subjects in the BBN for matching subjects across cameras, the MD is defined as:

$$\mathcal{M}_c = \frac{1}{v}\sum_{j=1}^{v}\sum_{i=1}^{u}[(\lambda_j - \mu'_{2,i})^T \Sigma_{2,i}^{-1}(\lambda_j - \mu'_{2,i})]\ p(i) \tag{6.9}$$

is used, where v is the number of pixels sampled from a subject's image. To compare the candidate matches in the comparison node of apparent colour modality, the accumulated information

$$M = \frac{1}{\sum_{i=0}^{q-1}\alpha_i}\sum_{i=0}^{q-1}\alpha_i \mathcal{M}_c(l-i) \tag{6.10}$$

is used, where q is the number of frames of accumulated evidence, l is the frame index and α is the weight to set more recent evidence with higher weights (as discussed in Section 4.4.2).

To determine the comparison result (i.e. similar or not similar) between each pair of subjects based on apparent colour, Equation (5.10) is used again. In this case, $\mathcal{X}_T^2$ is set to 5.99 corresponding to a 95% probability of finding the true feature for $n_f = 2$ (i.e. dimensionality of the colour vector).

To adjust the modality confidence of apparent colour during tracking, the confidence indicator is defined in terms of the segmentation status of the highest point of the subjects, defined in Section 5.2.2. This is because the position of the sub-image, where colour samples are taken, depends on the position of the highest point (as discussed in Section 6.2.1). When the position of the sub-image is not reliable, the apparent colour modality is also less reliable. On the other hand, the experiments performed (e.g. the 4^{th} test sequence pair in Table 7.1 in Section 7.2.3) indicate that the colour distribution can vary significantly when clothes reflects the illuminants. The distance between the means of the dominant Gaussian variables modelling the colour of a subject's images in two consecutive frames is used as the other indicator. This mean distance is defined as $D = \frac{1}{m}\sum_{i=1}^{m} d_i$, where m is the number of subjects in a camera image, and d is the distance on HS plane between the means of the dominant Gaussian variables in two consecutive frames. The state of this confidence indicator is defined as:

$$C_i(l) = \begin{cases} \text{high confidence} & ,\quad D \leq 0.02 \\ \text{medium confidence} & ,\quad 0.02 < D \leq 0.04 \\ \text{low confidence} & ,\quad D > 0.04 \end{cases} \tag{6.11}$$

where l is the frame index. When this distance is large, the confidence is set lower. This is because the reflection may result in a significant colour shift so that the estimated colour may be less reliable.

Moreover, $D = \frac{1}{a}\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m} d_{i,j}$ is used to indicate the confidence, where $a = \sum_{i=1}^{m-1} i$, m is the number of subjects in a camera image, and d is the distance on HS plane between means, μ, of the dominant Gaussian variables of each pair of subjects in an image. The state of this confidence indicator is defined as:

$$C_i(l) = \begin{cases} \text{high confidence} & , \quad D > 0.1 \\ \text{medium confidence} & , \quad 0.05 < D \leq 0.1 \\ \text{low confidence} & , \quad D \leq 0.05 \end{cases} \tag{6.12}$$

This value is used to make the system rely less on colour information when the colours between different subjects are similar (i.e. when the distance shorter).

6.4.2 Using Estimated Apparent Height for Correspondences

To match subjects in two camera images based on apparent height, the learnt mapping is used to estimate the apparent height of the corresponding subject across camera images. The goal is, for a subject in image I_1 with a VA, x_1, and apparent height, h_1, to search for the corresponding subject in image I_2. In doing so, the learnt mapping is used to estimate the apparent height, h', for each subject in I_2 with a VA, x_2, based on the observations (x_1, x_2, h_1). Even though the corresponding subject in I_2 is not known, the observed apparent height h of the corresponding subject in I_2 should ideally be equal to the estimated apparent height h' (assuming the learnt mapping can correctly estimate the apparent height). Therefore, the likelihood of a subject in I_2 with apparent height h, being the subject in I_1 can be evaluated based on the value of $\Delta h = h - h'$. Again, this difference is modelled as a Gaussian variable with a zero mean. Thus, the matching likelihood of a subject in I_2 with Δh is given by:

$$f(\Delta h) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(\Delta h)^2}{2\sigma^2}). \tag{6.13}$$

To compare the candidate matching subjects, the MD, defined as $\mathcal{M}_{ht} = \frac{(\Delta h)^2}{\sigma^2}$, is used in the comparison node of the BBN. Thus, the accumulated information of the apparent height used in the comparison node is:

$$M = \frac{1}{\sum_{i=0}^{q-1}\alpha_i}\sum_{i=0}^{q-1}\alpha_i \mathcal{M}_{ht}(l-i), \tag{6.14}$$

where q is the number of frames of accumulated evidences, l is the frame index and α_i is the weight used for setting more recent evidence with higher weights.

To determine the comparison result (i.e. similar or not similar) between each pair of subjects based on apparent height, Equation (5.10) is used again. In this case, $\mathcal{X}_T^2$ is set to 3.84 corresponding to a 95% probability of finding the true feature for $n_f = 1$ (i.e. dimensionality of the apparent height vector).

To indicate the modality confidence, the confidence indicator is defined in terms of the segmentation status of both the highest (defined in Equation (5.12)) and the lowest points of the subjects, which are used for computing apparent height. The segmentation status of the lowest point is defined as the mean distance, $D = \frac{1}{m}\sum_{j=1}^{m} d_{c,j}$ (pixels), of m subjects' lowest points in a camera image where d_c is the image distance between the lowest points (x, y) of a tracked subject in two consecutive frames. D is used as a confidence indicator to reduce the confidence when this distance is large. The state of this confidence indicator is defined as:

$$C_i(l) = \begin{cases} \text{high confidence} & , \quad D \leq 4.0 \\ \text{medium confidence} & , \quad 4.0 < D \leq 10.0 \\ \text{low confidence} & , \quad D > 10.0 \end{cases} \tag{6.15}$$

On the other hand, the mean difference of apparent heights between all m subjects in an image is also used as a confidence indicator. The mean difference is computed based on $D = \frac{1}{a}\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m} d_{i,j}$ (pixels), where $a = \sum_{i=1}^{m-1} i$ and d is the difference between the apparent heights of two subjects. When the mean difference of subjects' apparent heights is small, the modality confidence is reduced. The state of this confidence indicator is defined as:

$$C_i(l) = \begin{cases} \text{high confidence} & , \quad D > 20.0 \\ \text{medium confidence} & , \quad 10.0 < D \leq 20.0 \\ \text{low confidence} & , \quad D \leq 10.0 \end{cases} \tag{6.16}$$

6.4.3 Discussion

This learning approach can be generalised to different space-partitioned methods, where the mapping can be learnt with respect to these sub-spaces. It can also be generalised to other image features as long as the corresponding features in the two camera images hold the one-to-one relation. Figure 6.9 illustrates that the apparent width of a subject may not be applicable as a constraint for matching subjects across cameras due to its potential ambiguity. Two frames in two sequences captured from two cameras are shown. The top row shows the 30^{th} frame and the bottom the 130^{th}. The apparent widths of

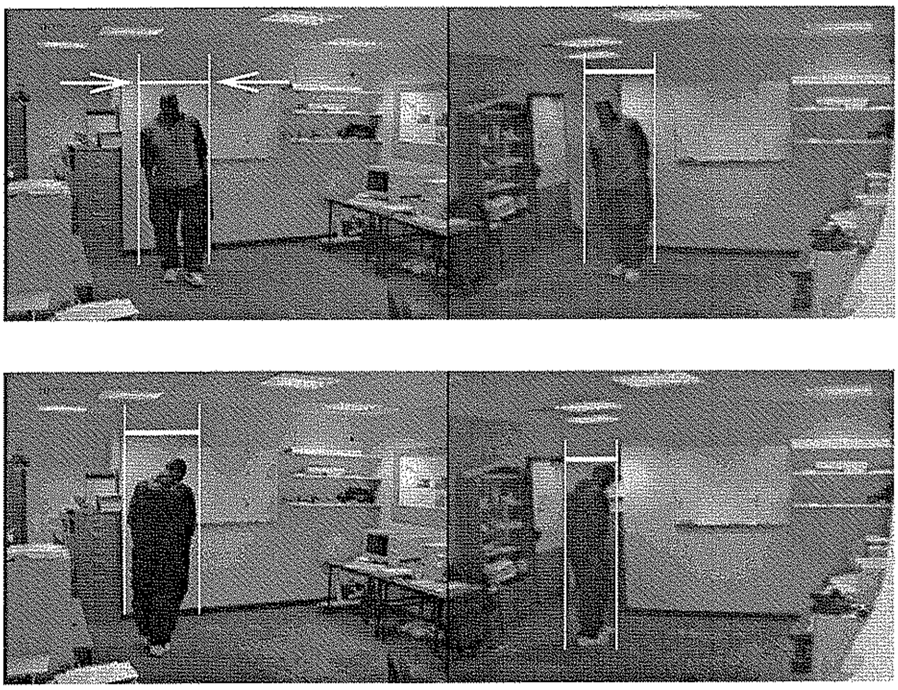


Figure 6.9: *An example demonstrating that the relation of apparent widths of a person in two camera images is not one-to-one.*

the person in the two frames from the left camera are virtually the same, but are very dissimilar to each other in the right camera. Thus, the apparent width of a person in two camera images does not hold the one-to-one relation.

In this approach, the mappings for different colours and heights are learnt, assuming that the mappings of a certain appearance value are the same for all world points in a cell. Although the learnt mapping of the apparent colour applies well to the scenario with the learnt illumination, it may not apply to un-learnt illumination conditions. On the other hand, the mapping of subjects' apparent heights between two camera images may not apply to conditions where background objects have been removed or added, since this can change the correlation of the apparent height between two camera images. However, this modality does apply to subjects with different heights and different poses since the mapping is learned for different apparent heights. It is also applicable even when the lower body parts of subjects are not visible in either or both camera images.

6.5 Summary

This chapter has described the recognition-based modalities, apparent colour and apparent height, used for matching subjects across cameras. As discussed in Chapter 1, the system views the same scene from very different viewing angles. There are significant image variations which can cause incorrect matches when directly using the subject appearance in one camera image to search for the corresponding subject in the other camera image. In order to compensate for the image variation in subject appearance between two camera images, SVR (Support Vector Regression) is used to learn the mapping in order to estimate subject appearance across cameras. Since subject appearances in two camera images are highly correlated with actual world positions, the mappings are learnt for different cells in the scene. The data collected in the comparison nodes and confidence indicator nodes for different modalities were explained. The collected visual evidence is fused by the BBN in order to probabilistically infer the subject correspondences between two camera images. The parameters used and the experimental results regarding the methodology for MCCT (Multiple Camera Cooperative Tracking) are given in the next chapter.

Chapter 7

Multi-Camera Cooperative Tracking

The tracking system proposed in this thesis has two tracking modes: SCT (Single Camera Tracking) and MCCT (Multiple Camera Cooperative Tracking). The experimental results of SCT were given in Chapter 3. This chapter demonstrates the experimental results of MCCT, which matches subjects images across cameras in order to track multiple people using two widely separated cameras cooperatively.

A Tracking System

The goal of the work presented in this thesis is to coordinate multiple un-calibrated cameras for the purpose of tracking multiple people. To fuse multi-camera data for tracking people, the system needs to match subjects between different cameras. This process of inter-camera subject correspondences allows the system to track people using multiple cameras cooperatively. The main difficulty of matching subjects across cameras lies in the correspondence process. For widely separated cameras, the correspondence is more challenging due to large image variations. In order to test our system developed for MCCT, two different types of cameras are deliberately used as image variations are larger due to different camera parameters. These two Charge-Coupled Device (CCD) cameras (a SGI digital camera and a SONY EVI-D31 camera) are connected to a SGI Octane workstation running the IRIX 6.5 operation system. The images were captured with an Octane video board with a frame rate of 25 Hz and handled off line on a SGI workstation using the clipped image frame format (400×300).

Experiments on MCCT

To perform feature correspondence for MCCT (see Figure 4.1), the system needs to perform two steps: preprocessing and matching. The preprocessing step includes two stages: (1) change detection and grouping and (2) feature extraction. The first stage, change detection and grouping, was illustrated in Section 3.2.1 (see Figures 3.5 and 3.6). The second stage is to extract the features from the subject image. The feature used for applying geometry-based modalities is the highest point of the subjects. Feature extraction of this point was demonstrated in Section 3.2.1 (see Figure 3.7). The features used for recognition-based modalities are apparent colour and apparent height of the subject. Extraction of these two appearance features was illustrated in Chapter 6 (see Figures 6.2 and 6.3).

The second step of performing MCCT is to match subjects across cameras. The task of matching subject's images across cameras is achieved by establishing feature correspondence based on multiple features. The following sections describe three experiments related to this matching task. The first experiment, described in Section 7.1, involves selection of Gaussian variable parameters for different modalities. These Gaussian variables are used to model the matching likelihood given the attribute disparity of features extracted from the subjects in two camera images. Section 7.2 describes the second experiment, which focuses on the estimation of subject appearance across two camera images for recognition-based modalities. As mentioned in Chapter 6, the central aim of mapping is to compensate for the image variation between two camera images in order to make inter-camera subject correspondences more reliable. The results of the estimated subject appearance based on the learnt mapping will show how close we are to this objective. Also, the robustness of appearance estimation using the learnt mapping between two cameras based on SVR is demonstrated through comparison with a second method. This second method was used to test the theory of mapping learning during research, and is an un-supervised learning method, Hierarchical Principal Component Analysis (HPCA). The third experiment, given in Section 7.3, investigates the application of Bayesain modality fusion for matching subjects across cameras. Since large image variations between camera images, as well as data uncertainty, can result in less reliable matches, the system fuses multiple modalities, deals with data uncertainty and captures correlation between modalities by using a BBN. An example of tracking multiple people using two widely separated cameras is illustrated. The identities of people are maintained by matching subjects in two camera images based on Bayesian modality fusion of multiple modalities. To highlight the strength of Bayesian modality fusion for combining multiple modalities, it is compared with a popular fusion method, often called *naive Bayes* method, which assumes all modalities are independent.

7.1 Experiments on Modelling the Matching Likelihood

To match subjects across camera images, different features from a subject's image are extracted and used. The Gaussian variables are used to model the matching likelihood for applying the modalities of homography, epipolar geometry and apparent height. The matching likelihood values between subjects are computed from the attribute disparities between the features of the two subjects in two cameras. These attribute disparities are assumed to be Gaussian distributions with means of zero. The parameters of the Gaussian variables are obtained from experiments described in the following. An example is also given to illustrate detection of matching ambiguity in MCCT (discussed in Section 4.4.3), based on the MD (Mahalanobis Distance), using the obtained Gaussian parameter of apparent height modality.

7.1.1 Homography

As explained in Section 5.2.2, to apply homography for matching subjects across cameras, the highest point of a subject in image I_1 is transferred to I_2 to compute the kinematic vector, $\mathbf{x}' = (x', y', \dot{x}', \dot{y}')$, in order to search for the corresponding subject in I_2. The matching likelihood of a subject with observed kinematic vector $\mathbf{x}$ in I_2 is given by a probability density function. For convenience, the function is given here:

$$f(\triangle\mathbf{x}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} exp(-\frac{1}{2}[(\triangle\mathbf{x})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\triangle\mathbf{x})]), \tag{7.1}$$

where $\triangle\mathbf{x} = (\mathbf{x} - \mathbf{x}')$.

Figures 7.1, 7.2, 7.3 and 7.4 show the experimental results of the measured and estimated kinematic vector of the subject in Figure 5.4 for 600 frames of a sequence pair. This sequence pair is taken from two cameras where the person walks around the office. Throughout the whole sequence, there is only one person moving around in a room. The highest points of the person in two camera images are extracted. The highest point of the subject in the right camera image is transferred to the left camera image. This transferred point is used to compute the estimated kinematic vector, $\mathbf{x}'$, for

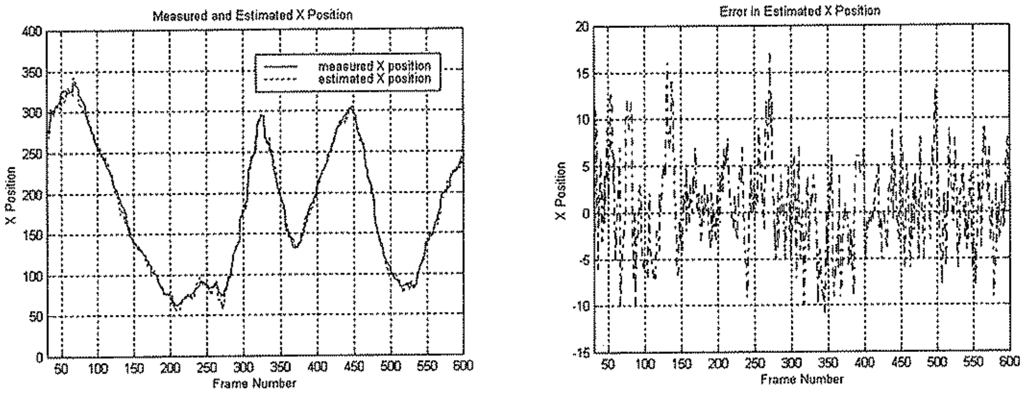


Figure 7.1: *Measured X position (pixels) and estimated X position (pixels) (based on homography) of a person's highest point in the left camera image for a sequence pair from two cameras.*

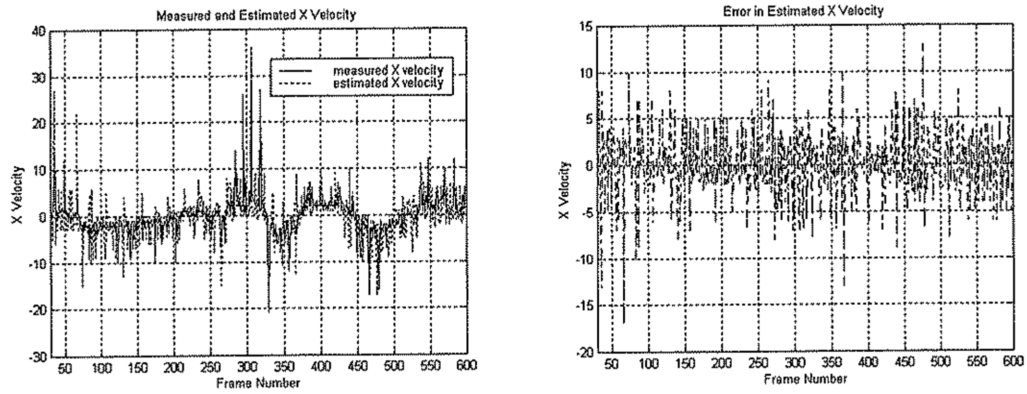


Figure 7.2: *Measured X velocity (pixels per frame) and estimated X velocity (pixels per frame) (based on homography) of a person's highest point in the left camera image for a sequence pair from two cameras.*

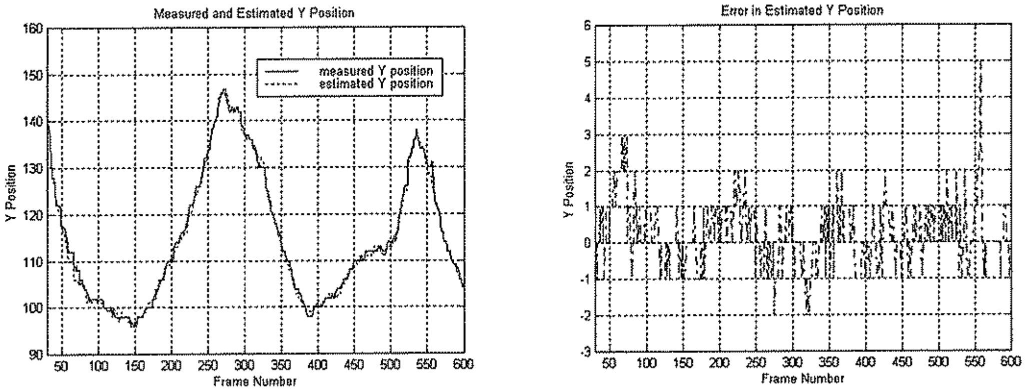


Figure 7.3: *Measured Y position (pixels) and estimated Y position (pixels) (based on homography) of a person's highest point in the left camera image for a sequence pair from two cameras.*

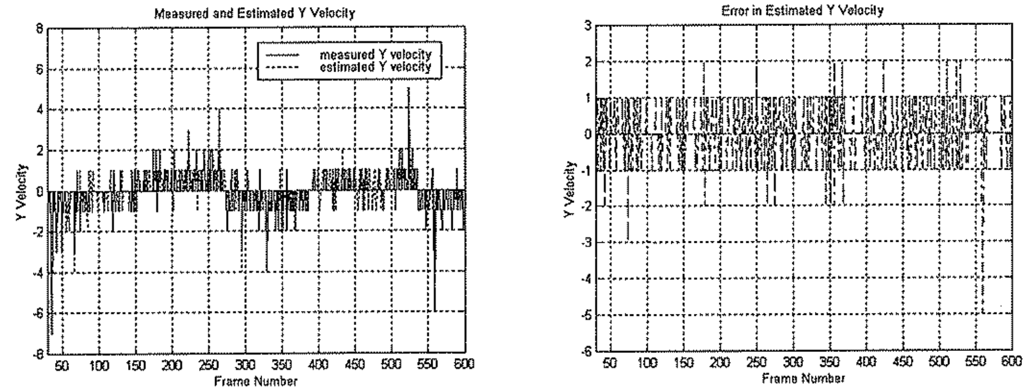


Figure 7.4: *Measured Y velocity (pixels per frame) and estimated Y velocity (pixels per frame) (based on homography) of a person's highest point in the left camera image for a sequence pair from two cameras.*

the subject in the left camera image. The highest point of the subject in the left camera image is used to compute the measured kinematic vector, $\mathbf{x}$. Ideally, the error value, $\triangle\mathbf{x} = (\mathbf{x} - \mathbf{x}')$, between the measured and estimate kinematic vector should be equal to zero. For the whole sequence, the mean values of error between the measured and estimated kinematic vector, $E[\triangle\mathbf{x}]$, are (0.636, 0.267, 0.005, 0) with standard deviations (4.823, 0.900, 3.561, 0.751). Finally, the covariance matrix, $\mathbf{\Sigma}$ is obtained as:

$$COV(\mathbf{x}) = \begin{pmatrix} 23.262 & -0.679 & 6.346 & -0.409 \\ - & 0.809 & -0.331 & 0.282 \\ - & - & 12.683 & -0.739 \\ - & - & - & 0.564 \end{pmatrix}. \tag{7.2}$$

These learnt parameters are then used in Equation (7.1) in order to compute the matching likelihood of two subjects in two camera images based on homography modality.

7.1.2 Epipolar Geometry

To apply epipolar geometry for matching subjects across cameras, the highest point is used as a feature point to represent the image position of a subject. The highest point of a subject in one camera image, I_2, is used to compute its associating epipolar line in the other camera image, I_1 (see Figure 5.9). The matching likelihood of a subject in I_1 (with a distance x between one's highest point and the epipolar line) being the subject in I_2 is given by the probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{x^2}{2\sigma^2}). \tag{7.3}$$

Figure 7.5 shows the distance between the epipolar line and the highest point of subject 2 (in the left camera image in Figure 5.9) for 600 frames of a sequence pair from two cameras. The epipolar line is computed using the highest point of subject B in the right camera image. The sudden change in the distance, at the 40^{th} frame, is due to incorrect position of the extracted highest point. The mean distance value for the whole sequence is -0.249 (pixels), with a standard deviation $\sigma = 1.810$.

7.1.3 Apparent Height

To match subjects in two camera images based on apparent height, the mapping of apparent heights is learnt using SVR (see Section 6.3.2). Based on the learnt mapping, the apparent height of a subject in image I_1 is used to estimate the apparent height,

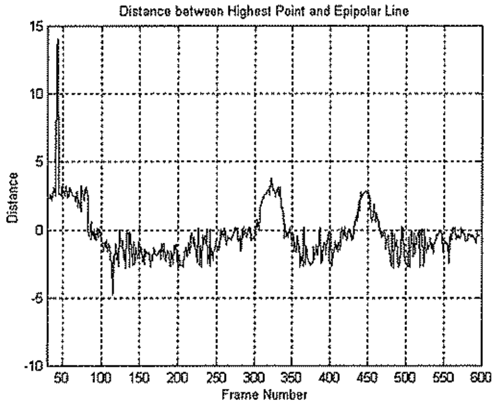


Figure 7.5: *Distance between the highest point of a subject and the epipolar line.*

h', to search for the corresponding subject in image I_2. The likelihood of a subject in I_2 with observed apparent height, h, being the subject in I_1 is evaluated based on the value of $\triangle h = (h - h')$. The matching likelihood of a subject in I_2 with $\triangle h$ is given by:

$$f(\triangle h) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(\triangle h)^2}{2\sigma^2}). \tag{7.4}$$

Figure 7.6 shows the observed and estimated heights of the subject in the left camera image of Figure 6.7 for 380 frames of a sequence pair from two cameras. The estimated height is computed based on the observed apparent height of the subject in the right camera image. The mean value of the estimation error for the whole sequence is 1.564 (pixels), with a standard deviation $\sigma = 11.659$. It is important to point out that compared to the standard deviation, the mean estimation errors for all three modalities are relatively small and close to zero. Thus, the assumption that using Gaussian variables with zero means to model the matching likelihood for these modalities based on the attribute disparity is largely valid.

7.1.4 Matching Ambiguity in MCCT

To illustrate an example of detecting matching ambiguity in MCCT (see Section 4.4.3), the modality of apparent height is used. Figure 7.7 shows the computed MD, $\mathcal{M}_{ht} = \frac{(\triangle h)^2}{\sigma^2}$, for apparent height of the subject in the left camera image in Figure 6.7 for 380 frames of a sequence pair from two cameras. The error $\triangle h$ is computed by $(h - h')$ where h and h' are the observed and estimated apparent heights in the left camera image, respectively (see Section 6.4.2). The estimated apparent height, h', of the subject in the left camera image is computed from the apparent height of the subject in the right camera estimate the apparent height. This may result from the fact that the cell volume (see Figure 5.13), used to learn the mapping of apparent height between two camera images (as discussed in Section 6.4.2), is too large. So, the assumption made in appearance estimation that the mappings of apparent height for the whole cell are the same is no longer valid when incorrect ambiguity detection occurs.

7.2 Experiments on Appearance Mapping across Cameras

To apply recognition-based modalities, the mapping learnt in the training phase is used to estimate subject appearance for matching subjects across cameras during tracking (see Figure 6.1). Firstly, the experimental training phase results are illustrated, before the results of estimation of subject appearance are given. Also, the estimation of subject appearance based on SVR is compared to the estimation based on HPCA.

7.2.1 Training Phase

As discussed in Section 6.3.2, in order to compensate for colour shift between two camera images, the apparent colour of a subject in one camera image is used to estimate the mean position of the colour distribution, (μ_h, μ_s), of the corresponding subject image in the other camera. Two SVR-based apparent colour mappings (one for μ_h and the other for μ_s) are constructed to estimate the apparent colour from one camera image to the other camera image and another one mapping constructed for the mapping of apparent height. In the experiments, a Gaussian kernel was used to build the SVR function (see Appendix F). It has been found that the kernel usually provides an acceptable performance when its parameter $2\delta = 1$ (see Equation (F.7)) and the input patterns were normalised to unit vectors. The regularisation penalty factor, C (see Equation (F.5)), is set to 1000. The tolerance coefficient ε defined in the loss function (defined in Equation (F.8) of Appendix F) for mapping of μ_h, μ_s and apparent height were set to 3° and 0.06 and 3 pixels, respectively.

Figures 7.8, 7.9, and 7.10 show the number of support vectors obtained and the training times for different numbers of sequence pairs the system was trained with. Each sequence pair was recorded with a single person in both camera images and with 550 frames. Different sequence pairs are recorded with different people at different heights and wearing different coloured clothes. Notice the *smooth* relationship between the number of support vectors and the training time with respect to the number of sequence

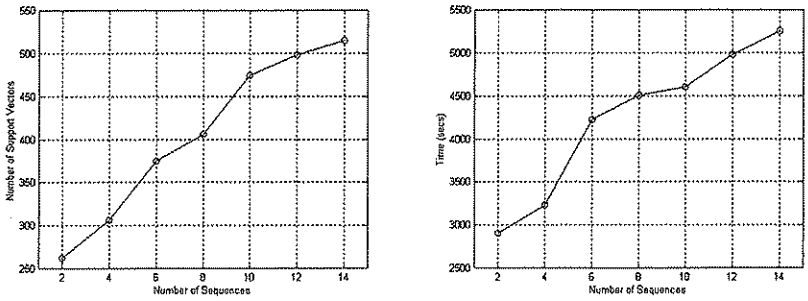


Figure 7.8: *Number of support vectors after training and training time of hue mapping.*

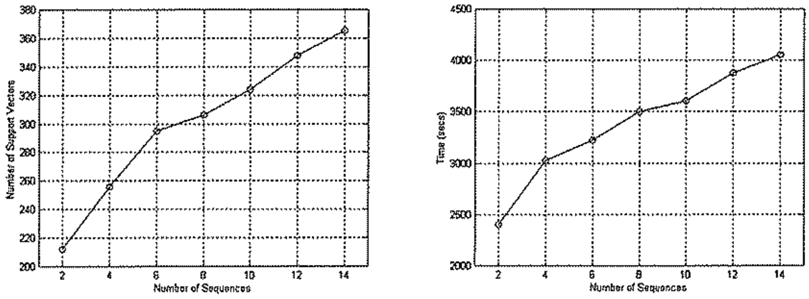


Figure 7.9: *Number of support vectors after training and training time of saturation mapping.*

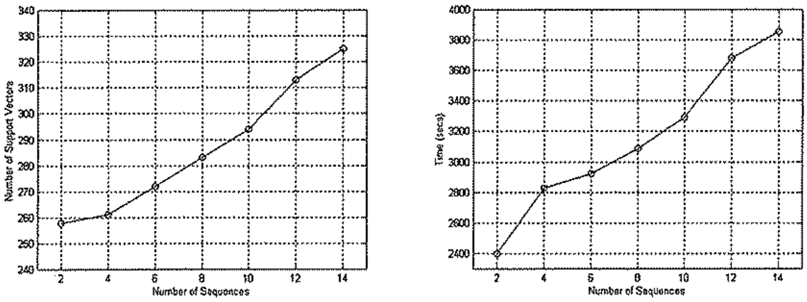


Figure 7.10: *Number of support vectors after training and training time of apparent height mapping.*

pairs. In general, the number of support vectors and the training time increase with the number of sequence pairs. However, the increasing rate of both number of support vectors and the training time decrease as the number of sequence pairs increases.

7.2.2 Estimation of Subject Appearance

In the experiment, it was found that the mapping trained with 14 different sequence pairs provided acceptable performance and was used for estimation of subject appearance across cameras. Figure 7.11 shows two examples of the sequence pair that used for training. The top row shows two images from the sequence pair with the highest person among these sequences and the bottom row shows two images with the shortest person. In each pair of these sequences, a person walks around in the room. The joint colour features and joint height features are extracted from two camera image for learning the mapping of joint features. Although more sequences can be used , it will increases the number of SVs required to construct the mapping function. To investigate the success of estimation of subject apparent colour across camera images, a pair of test sequence captured from two cameras are used. The results obtained from more test sequence pairs are given later, when a comparison is made with estimation using HPCA. In this pair of test sequences, a person walks around the office. The colour samples were taken from the person's clothes (see Figure 7.12). The mean of colour samples at each frame in two cameras, and the estimated colour mean at each frame are shown in Figure 7.13. The left graph shows the observed colour sample means in two cameras which is also shown in Figure 6.4 (in Figure 6.4, it can be seen that the apparent colour is highly correlated to the world position). The right graph shows the estimated colour means (using the learnt mapping based on SVR) together with the observed colour sample means. The estimated colours for the left camera are in general closer to the observed colours in the left camera than the observed colours in the right camera.

The central aim of appearance estimation is to compensate for image variation between two camera images so as to make inter-camera subject correspondences more reliable. This example (Figure 7.13) provides an qualitative measure of the accuracy of estimated apparent colour across cameras. In order to compare estimated colour with observed colour, the mean of all colour means for all frames was computed. The distance between the means of all the estimated and observed colour in the left camera is 14% of the distance between the means of observed colours in two camera images. This im-

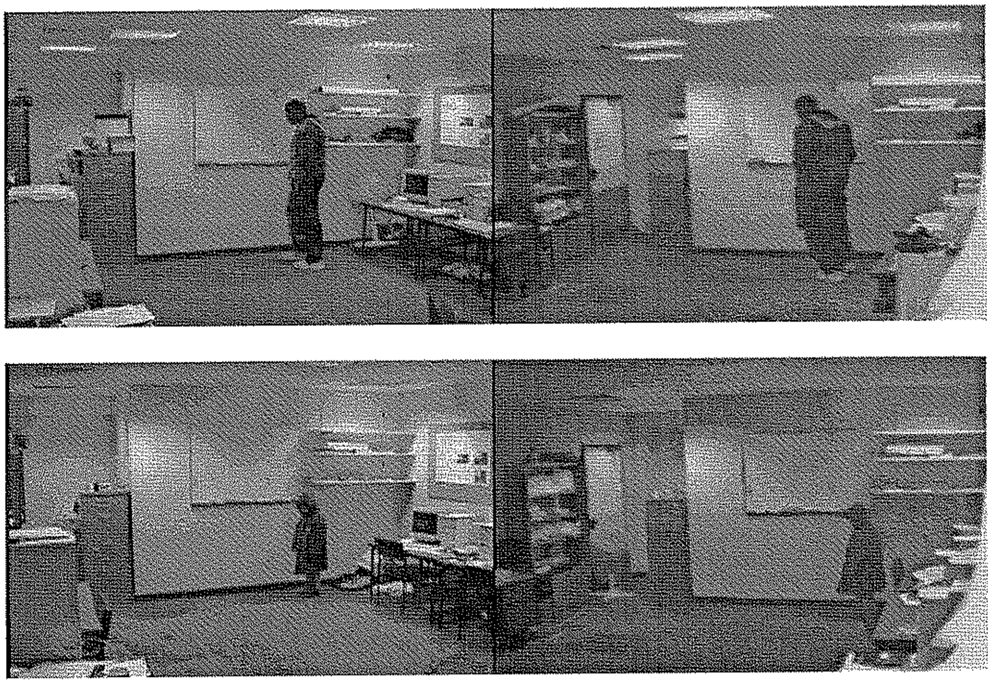


Figure 7.11: *Two example of the training sequence pairs.*

plies that on average, the estimated colour for the left camera is closer to the observed colour in the left than the observed colour in the right camera. Therefore, using the estimated colour to search for the corresponding subject across cameras is more reliable than directly using the observed colour (i.e. without estimation) as a model. On the other hand, the estimation of colour means corresponding to a person in a certain cell in the scene is also shown in Figure 7.14. The left graph shows the observed colour sample means from a person's clothes in two camera images which is also shown in Figure 6.5, (In Figure 6.5 it is used to show that the colour is highly correlated to the world position in the scene). The right graph shows the estimation results (using the learnt mapping based on SVR).

To illustrate the results of estimation of subject apparent height across cameras, a pair of test sequences are used where a person walks around the office and the apparent heights are extracted from two camera images. Figure 7.15 shows the observed apparent height in the two camera images and the estimated apparent height for the left camera image. The apparent height of a person in the two camera images is also shown in Figure 6.7 (in Figure 6.7, it can be seen that the apparent height is highly correlated to the world position). The apparent height in the right camera image drops significantly when the lower part of the subject is not visible due to partial occlusion by the image

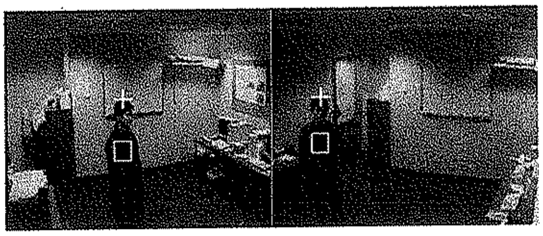


Figure 7.12: *Two images of the test sequence pair captured from two cameras.*

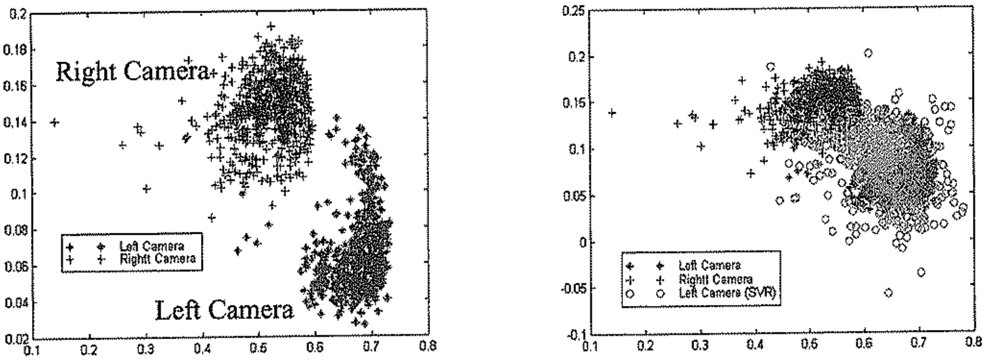


Figure 7.13: *Estimation of apparent colour across cameras over 400 frames. (Note that for convenience the polar coordinates of hue and saturation are drawn in Cartesian coordinates)*

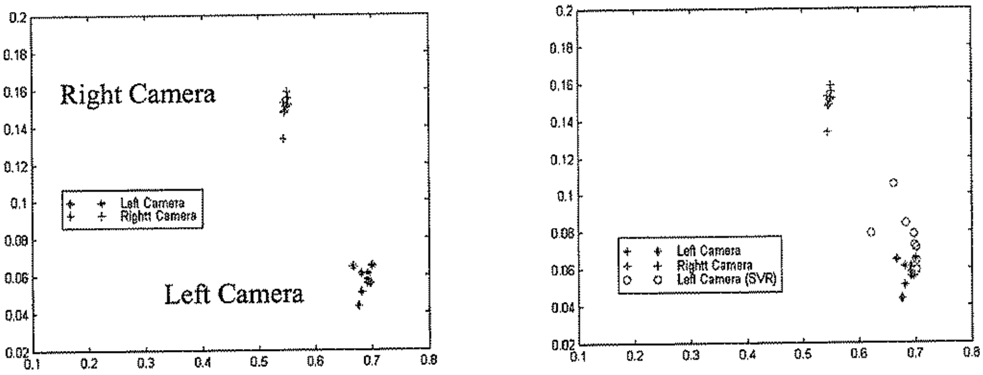


Figure 7.14: *Estimation of apparent colour across cameras for a certain cell. (Note that for convenience the polar coordinates of hue and saturation are drawn in Cartesian coordinates)*

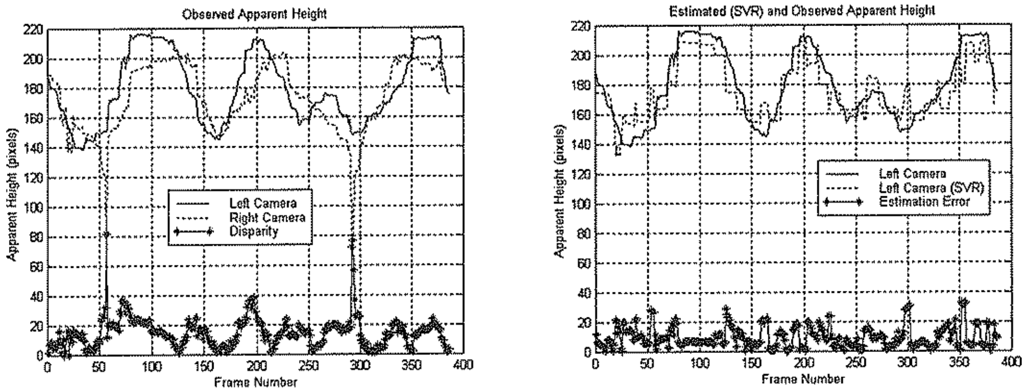


Figure 7.15: *An example of estimation of apparent height across cameras over 380 frames.*

boundary and objects in the room. The estimated apparent height in the left camera is based on the observed apparent height in the right. Note that the estimated height for the left camera is not affected by the sudden change of the observed apparent height in the right camera. Over the whole sequence, the absolute estimate error, $|h_1 - h'_1|$, is about 35.17% of the absolute height disparity, $|h_1 - h_2|$, between the two cameras, where h_i is the measured apparent height in image I_i and h'_1 is the estimated apparent height based on h_2.

7.2.3 Comparative Evaluation

To highlight the strength of using SVR for learning the mapping between subject appearance in two camera images, the estimation results are compared to results obtained using HPCA [60], which we used to test mapping learning during the research [28]. The training examples are formed as $(x_1, x_2, \mathbf{G}_1, \mathbf{G}_2)$ (i.e. joint colour feature, see Section 6.2.1) and (x_1, x_2, h_1, h_2) (i.e. joint height feature). The steps to perform HPCA for estimating subject appearance across camera images are as follows:

1. Perform PCA on all the training examples and transform all training examples to the PCA space, called the *parameter space.* The eigenvectors found are referred to as *global eigenvectors* of the parameter space.

2. Perform k-means cluster on the transformed data to obtain k clusters.

3. Perform PCA on each cluster to obtain the eigen-vectors, referred to as *local eigen-*

vectors, for representing each cluster.

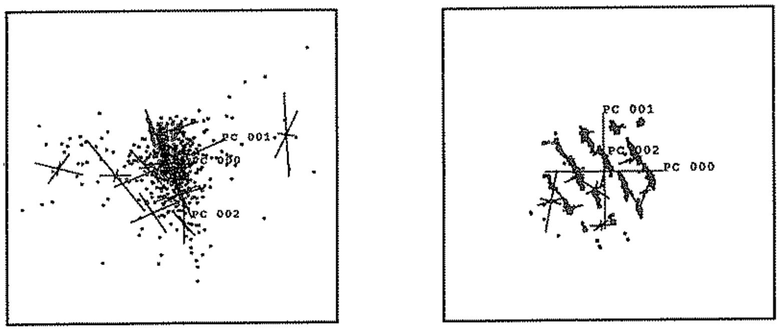
4. For an observation, $(x_1, x_2, \mathbf{G}_1)$ and (x_1, x_2, h_1), the missing part of the example (i.e. $\mathbf{G}_2$ and h_2, the appearance of the corresponding subject in the other camera image) is initially replaced with the most recent observation.

5. Project this synthesised vector into the parameter space and constrain this projected point to the nearest cluster in order to find the *most probable point* in the learnt distribution in the parameter space.

Figure 7.16 illustrates the learnt distribution of the parameter space using the 14 sequence pairs that were also used to train the SVR mappings. Figure 7.17 illustrates the three largest global eigenvectors which joint feature vectors are projected to. Training example vectors are shown with the local eigenvectors of different clusters. The missing part of the example can then be found from the most probable point. Tables 7.1 and 7.2 illustrate the estimation results for 5 test sequence pairs, each of 550 frames, based on SVR and HPCA, where the colour sample in each frame is modelled with 2 Gaussian variables. The overall mean of the dominant colour models in all frames of each sequence pair is listed in Table 7.1, together with the standard deviation. Table 7.2 presents the absolute relative disparity (between the observed apparent heights in two camera images) and absolute estimation relative error (for SVR and HPCA) of the apparent height. The absolute relative disparity is computed by $\frac{|h_1 - h_2|}{h_1}$, and the absolute estimation relative error is computed by $\frac{|h_1 - h_1'|}{h_1}$, where h_1 and h_2 are the observed heights of a person in the left and right camera respectively, and h_1' is the estimated apparent height for the subject in the left camera based on h_2 using SVR and HPCA. These experimental results indicate that:

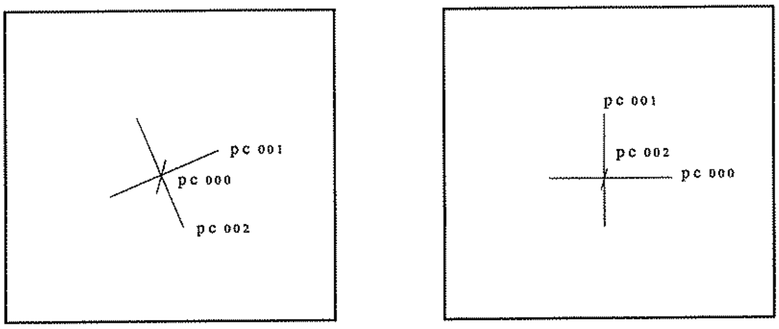
- Based on subject appearance in the camera image I_i, both methods can obtain an estimated appearance for the corresponding subject in I_j which is closer to the observed appearance of the subject in I_j than the observed appearance of the subject in I_i.

- SVR outperforms HPCA for both apparent colour and apparent height in the experiments performed. For all 5 test sequence pairs, the distance between the means of the estimated colour, based on SVR, and observed colour in the left camera is 19.28% of the distance between the means of observed colours in two camera images, while for HPCA this value is 56.66%. For apparent height over 5



Joint apparent colour Joint apparent height

Figure 7.16: *An illustration of the projections of joint feature vectors to the three largest global eigenvectors.*



Joint apparent colour Joint apparent height

Figure 7.17: *An illustration of the three largest global eigenvectors which joint feature vectors are projected to.*

test sequence pairs, the absolute estimation relative error, for SVR, is 15.78% of the absolute relative disparity, while it is 18.16% for HPCA.

- For the 4^{th} test sequence pair (see Table 7.1), the high deviation in the values of hue results from strong reflections from the clothes, which cause the colour distribution to vary significantly.

Based on the analysis above, the system can make subject correspondences more reliable based on SVR mapping than directly applying the appearance in I_i as a model to search for the corresponding subject in I_j. The experiments also suggest that the colour modality is less reliable when the colour distribution varies significantly. Therefore, when the colour shifts significantly between consecutive frames, the modality confidence should be reduced, as discussed in Section 6.4.1.

Table 7.1: *Estimated apparent colours for 5 sequence pairs.*

	Right mean (std)	Left mean (std)	Left (SVR) mean (std)	Left (HPCA) mean (std)
hue (1)	16.23° (5.01°)	7.16° (5.04°)	7.72° (5.16°)	10.74° (8.33°)
saturation(1)	0.54 (0.05)	0.66 (0.06)	0.64 (0.06)	0.60 (0.09)
hue (2)	27.26° (10.53°)	11.49° (6.43°)	13.81° (5.87°)	18.61° (13.92°)
saturation(2)	0.31 (0.06)	0.51 (0.06)	0.50 (0.08)	0.50 (0.08)
hue (3)	52.66° (8.73°)	44.51° (1.63°)	41.56° (5.51°)	40.09° (6.64°)
saturation(3)	0.32 (0.04)	0.42 (0.06)	0.39 (0.05)	0.41 (0.03)
hue (4)	164.36° (86.37°)	130.96° (37.02°)	124.95° (36.45°)	163.47° (51.7°)
saturation(4)	0.45 (0.05)	0.36 (0.07)	0.33 (0.08)	0.39 (0.03)
hue (5)	215.62° (3.24°)	218.78° (2.53°)	216.56° (2.48°)	229.82° (4.96°)
saturation(5)	0.12 (0.03)	0.06 (0.03)	0.06 (0.05)	0.22 (0.08)

Table 7.2: *Absolute estimation relative error in apparent height for 5 sequence pairs.*

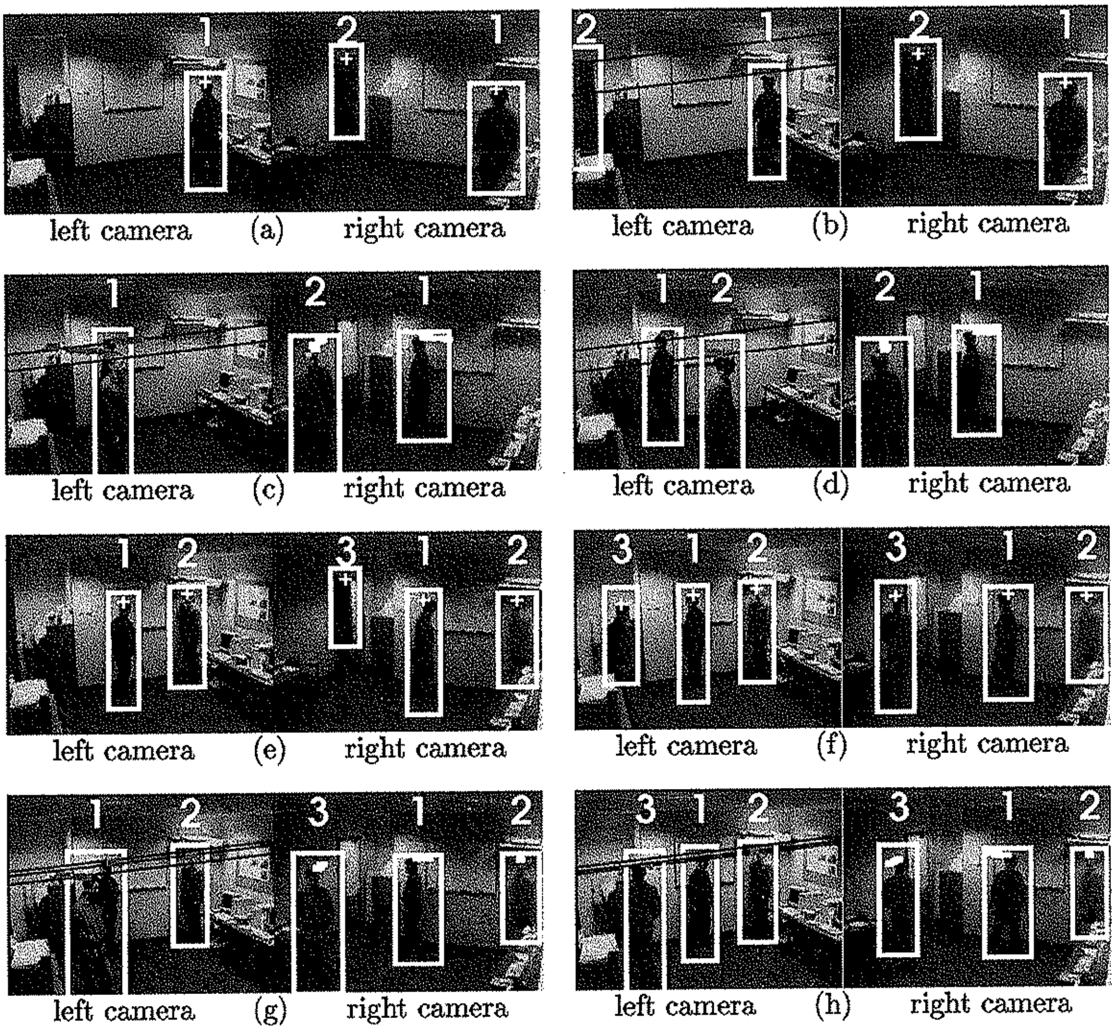
	Disparity	SVR	HPCA
sequence 1 (550 frames)	43.76 %	5.51 %	7.05 %
sequence 2 (550 frames)	44.32 %	7.34 %	8.08 %
sequence 3 (550 frames)	46.84 %	6.28 %	7.11 %
sequence 4 (550 frames)	31.75 %	8.41 %	8.65 %
sequence 5 (550 frames)	38.47 %	4.84 %	6.36 %

7.3 Experiments on Matching People across Cameras

7.3.1 Tracking People Using Two Cameras Cooperatively

Figure 7.18 demonstrates Multiple Camera Cooperative Tracking (MCCT) by showing a sequence pair from two cameras with three people interacting with one another. The system matches subject images across cameras in order to track people with the assigned identities over time. In order to test the system, all three people wear red clothes so that the algorithm cannot distinguish them based on colour alone. The label on top of a bounding box is the identity assigned by the system when the person first appears in either camera image. The white cross is the highest point of a subject. The system maintained identities consistently by performing MCCT throughout the whole sequence based on Bayesian modality fusion, even when occlusion is present in a camera image.

Note that if the computational power is limited, the system can perform MCCT only when matching ambiguity (i.e. occlusion) occurs in a camera in order to obtain subject identities from the other camera.



left camera (a) right camera

left camera (b) right camera

left camera (c) right camera

left camera (d) right camera

left camera (e) right camera

left camera (f) right camera

left camera (g) right camera

left camera (h) right camera

Figure 7.18: *A tracking example.*

To illustrate the working of the Bayesian modality fusion approach, a section of the sequences is highlighted when person 1 is visible in both camera images, and as person 2 enters the room, imaged by the right camera but not the left (Figure 7.18(a)). As person 2 enters the left FOV, both people are in the overlapping FOVs (Figure 7.18(b)), and from MCCT the system obtains the identity (for this newly detected subject in the left camera) from the right camera, assuming the subjects in two camera images are both in the overlapping FOVs. Based on the highest points of two subjects in the right camera image (I_2), the epipolar line (black) is computed and used to search for corresponding subjects in the left camera image (I_1). The highest point of person 1 is transferred to

I_1 (blue dots on top of person 1 for most recent 5 frames) based on the on-line learnt homography (induced by a virtual plane, see Figure 5.2). This transferred point is used for comparing with the observed kinematic vector (defined in Section 5.2.2) of two subjects in I_1. The system cannot use homography induced by person 2 for matching, since person 2 had just entered the room and his related homography was yet to be established. It is important to point out that even though the homography related to person 2 had not been established, BBNs can still infer the subject correspondences based on partial information (i.e. the evidence based on the homography related to person 1 and other modalities). It can also be seen that the highest point of person 2 is incorrectly segmented in the right camera image due to noise, as shown in Figure 7.18(b). Although the information is incomplete and less reliable, the BBN can still effectively collect the evidence and make a correct match to pass subject identity across cameras.

After entering, person 2 continues to walk towards the room centre and person 1 keeps walking towards the door. These two subjects meet in I_1 and are segmented as one single blob, but not in I_2 (Figure 7.18(c)). The system interprets that I_1 is ambiguous and relies on the tracking results from I_2 to disambiguate. The blue dots in I_1 are the transferred points from the highest points (white dots) of two subjects in I_2 based on the stored established homography. From the results of modality fusion, the merged blob in I_1 is matched to and interpreted as person 1 due to the top point of this blob corresponding to person 1.

Occlusion Resolved by MCCT

In this occlusion case, the system maintained the identities after occlusion by cooperatively using two cameras. When the merged blob splits into two blobs, the system detects that the number of blobs changes Figure 7.18(d). From the matching in MCCT, the system passes identities of two people from I_2 to I_1. Person 2 keeps walking to the right corner and person 1 turns and faces person 2. At this moment, another person enters the room and is assigned a new identity (i.e. person 3 in I_2, see Figure 7.18(e)). Person 1 then turns around and walks toward person 3 (Figure 7.18(f)). Similar to Figure 7.18(c), occlusion occurs in I_1 in Figure 7.18(g), but here the two people change direction during occlusion. To resolve this occlusion by MCCT, the homography modality is more reliable than the other modalities. It can be seen that in Figure 7.18(h), the transferred points in I_1 can be reliably used to search for the corresponding people in I_1 (based on the kinematic vector). The epipolar geometry is less reliable because the epipolar lines are close (see Section 5.3.2).

Kalman Filtering Failure

Note that tracking with a single camera based on Kalman filtering can resolve the ambiguity in the event of occlusion as shown in Figure 7.18(c), but cannot maintain correct identities for occlusion with direction change, as in Figure 7.18(g). Figure 7.19 illustrates the tracking failure with a single camera based on motion continuity using a Kalman filter for the latter event shown in Figure 7.18(g). It shows the measured and the predicted positions of the blob centroids of person 1 and person 3 in I_1. Occlusion is present during frames 343-395, and some example images taken before, during and after the occlusion event are shown in Figures 7.18(f-h). During occlusion, the position estimation is based on a constant velocity assumption and the acceleration is not used because it is unreliable. The Kalman filters can follow people before occlusion, but fail to estimate correct positions of people after occlusion where people change walking directions.

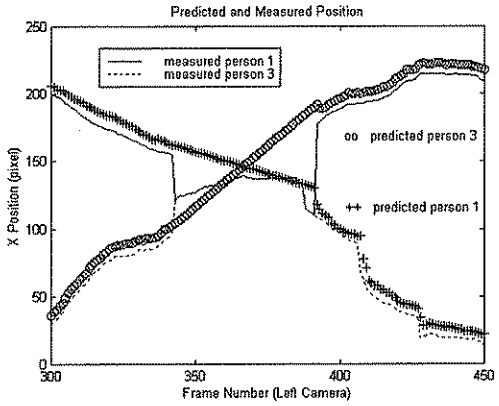


Figure 7.19: *The measured and predicted (using a Kalman filter) blob centroids of persons 1 and 3 in the left camera image of the tracking example (Figure 7.18).*

The Bayesian Belief Network

Network structure: Figure 7.20 shows the BBN used for the tracking example described above. It only shows the structure in a MU (Matching Unit, see Figures 4.9 and 4.10) for apparent height modality. The structures and parameters of the network for the other modalities use the same MUs, as described in Section 4.4.1. Since there are three people in the tracking example, the BBN is designed with $m = 3$, where m is the maximum number of subjects in two camera images. Thus, there are 3 comparison nodes

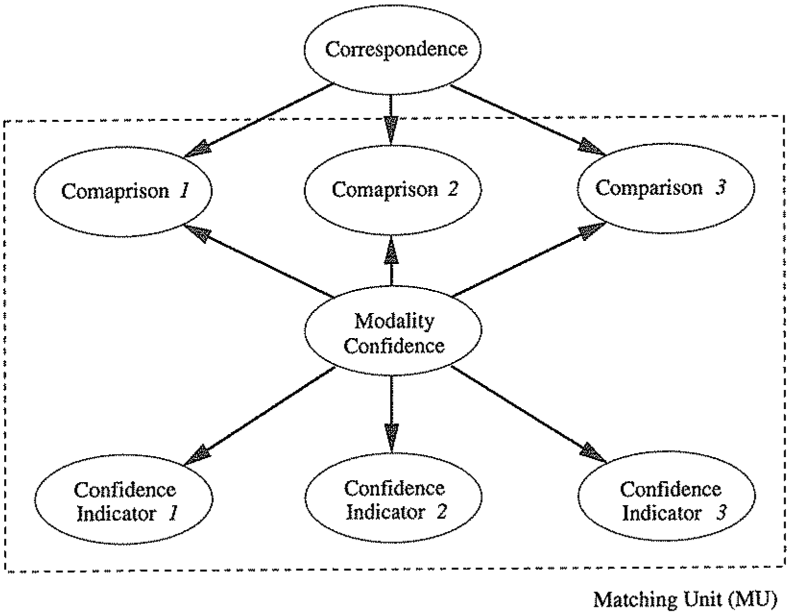


Figure 7.20: *The BBN (Bayesian Belief Network) for inferring the correspondence of subjects between two camera images based on the modality of apparent height.*

in the MU. Note that once the network has been designed, it can only handle the tracking scenario where the number of people are equal or less than m. Both the structure and the CPTs of the BBN can not be changed dynamically. If the computational power is available, a large network could be designed to handle more people. To indicate the modality confidence for apparent height (as described in Section 6.4.2), three confidence indicators are used, represented as three confidence indicator nodes. These indicators are the segmentation status of both the highest and the lowest points of the subjects, and the mean difference of m subjects' apparent heights.

States of variables: Table 7.3 lists the states of the variable represented by the correspondence node. For m subjects in each of the two images, there could be $m!$ possible assignment combinations. Thus, in total there are 6 assignment combinations, A_1-A_6. Each state shown in the *state table* corresponds to an assignment combination. Each assignment combination is defined as a union of 3 matches. Each match assigns one of the 3 subjects (S_1, S_2 and S_3) in the right camera to one of the 3 subjects (S_A, S_B and S_C) in the left camera.

In the network, each comparison node compares one of the 3 subjects in the left camera image with all 3 subjects in the right image. Table 7.4 shows the states of the variable represented by a comparison node which compares S_A with S_1, S_2 and S_3. Each

Table 7.3: *States of the variable represented by the correspondence node.*

States	Combination of Assignments
A_1	$S_A \leftarrow S_1$, $S_B \leftarrow S_2$, $S_C \leftarrow S_3$
A_2	$S_A \leftarrow S_1$, $S_B \leftarrow S_3$, $S_C \leftarrow S_2$
A_3	$S_A \leftarrow S_3$, $S_B \leftarrow S_1$, $S_C \leftarrow S_2$
A_4	$S_A \leftarrow S_2$, $S_B \leftarrow S_1$, $S_C \leftarrow S_3$
A_5	$S_A \leftarrow S_2$, $S_B \leftarrow S_3$, $S_C \leftarrow S_1$
A_6	$S_A \leftarrow S_3$, $S_B \leftarrow S_2$, $S_C \leftarrow S_1$

Table 7.4: *States of the variable represented by the comparison node.*

States	S_A and S_1	S_A and S_2	S_A and S_3
0	not similar	not similar	not similar
1	not similar	not similar	similar
2	not similar	similar	not similar
3	not similar	similar	similar
4	similar	not similar	not similar
5	similar	not similar	similar
6	similar	similar	not similar
7	similar	similar	similar

state encodes the comparison results with the 3 subjects. Note that the comparison results encoded in a state allow multiple hypotheses (i.e. multiple subjects in image I_j can be equally similar to a subject in I_i). The variables represented by all other comparison nodes have the same states table. Table 7.5 shows the thresholds (i.e. $\mathcal{X}_T^2$) used to determine the states (similar or not similar) of comparison results for different modalities.

On the other hand, the variable represented by the modality confidence node has only two states, i.e. high and low confidence. All the variables represented by confidence indicator nodes have three states, i.e. high, medium and low confidence. Table 7.6 shows the thresholds used to determine the states (high, medium or low confidence) for different confidence indicators.

Conditional probability tables of variables: As described in Section 4.3.3, each variable, V_i, is represented by a node, and has a set of conditional probabilities, $P(V_i|\Pi_{V_i})$, which is a function of its parent nodes Π_{V_i}. For the node with no parent node, this $P(V_i|\Pi_{V_i})$ degenerates to the prior $P(V_i)$. In the BBN used for fusing multiple modalities for MCCT (see Figure 4.9), the nodes have parent nodes are comparison and con-

Table 7.5: *Thresholds used for determining the states of the comparison results for different modalities.*

Modality	$\mathcal{X}_T^2$
Homography	9.49
Epipolar geometry	3.84
Apparent colour	5.99
Apparent height	3.84

Table 7.6: *Thresholds used for determining the states of the different confidence indicators. (D=distance).*

Confidence indicator	High	Medium	Low
Segmentation of highest points	$D \leq 2.0$	$2.0 < D \leq 4.0$	$D > 4.0$
Mean D between highest points	$D > 80.0$	$50.0 < D \leq 80.0$	$D \leq 50.0$
Mean D between epipolar lines	$D > 20.0$	$10.0 < D \leq 20.0$	$D \leq 10.0$
Mean D between subjects and landmarks	$D > 5.0$	$2.0 < D \leq 5.0$	$D \leq 2.0$
Mean D between Gaussians in two frames	$D \leq 0.02$	$0.02 < D \leq 0.04$	$D > 0.04$
Mean D between Gaussians of all subjects	$D > 0.1$	$0.05 < D \leq 0.1$	$D \leq 0.05$
Segmentation of lowest points	$D \leq 4.0$	$4.0 < D \leq 10.0$	$D > 10.0$
Mean D between apparent heights	$D > 20.0$	$10.0 < D \leq 20.0$	$D \leq 10.0$

fidence indicator nodes. Table 7.7 illustrates the CPT (Conditional Probability Table) of the comparison node which compares S_A with S_1, S_2 and S_3. It lists the belief of each state in the comparison node. This belief describes the probability distribution over the states, given the states of the parents nodes, i.e. the correspondence node and modality confidence node. For example, given the state A_1 in correspondence node and low confidence in modality confidence node, the probability distribution over the states of this comparison node are (4%, 4%, 4%, 4%, 40%, 20%, 20%, 4%). The state 4 of the comparison node corresponds to the comparison results of S_A similar to S_1 but not similar to either S_2 or S_3 (see Table 7.4). The reason for state 4 with the highest probability (i.e. 40%) is that this evidence (i.e. state 4) strongly supports state A_1 (i.e. $S_A \leftarrow S_1$) in correspondence node, compared to other states of comparison node. For example, state 5 corresponds to the comparison results of S_A similar to S_1 and S_3, but not similar to S_2 (see Table 7.4). State 5 equally supports $S_A \leftarrow S_1$ and $S_A \leftarrow S_3$. Thus, the probability is half that of state 4 and is set to 20%. This is because state 4 corresponds to the comparison results of S_A similar to S_1, but not similar to S_2 and S_3. All other comparison nodes have similar CPTs, but based on a similar principle.

If all the modalities give conflicting results, it is possible that the inferred probability distribution over $m!$ possible assignment combinations are equal. For example, $S_A \leftarrow S_1$

Table 7.7: *Conditional probability table of a comparison node given the states of its parent nodes. (M.C. = Modality Confidence).*

Correspondence	M. C.	0	1	2	3	4	5	6	7
A_1	low	4%	4%	4%	4%	40%	20%	20%	4%
A_1	high	2%	2%	2%	2%	60%	15%	15%	2%
A_2	low	4%	4%	4%	4%	40%	20%	20%	4%
A_2	high	2%	2%	2%	2%	60%	15%	15%	2%
A_3	low	4%	40%	4%	20%	4%	20%	4%	4%
A_3	high	2%	60%	2%	15%	2%	15%	2%	2%
A_4	low	4%	4%	40%	20%	4%	4%	20%	4%
A_4	high	2%	2%	60%	15%	2%	2%	15%	2%
A_5	low	4%	4%	40%	20%	4%	4%	20%	4%
A_5	high	2%	2%	60%	15%	2%	2%	15%	2%
A_6	low	4%	40%	4%	20%	4%	20%	4%	4%
A_6	high	2%	60%	2%	15%	2%	15%	2%	2%

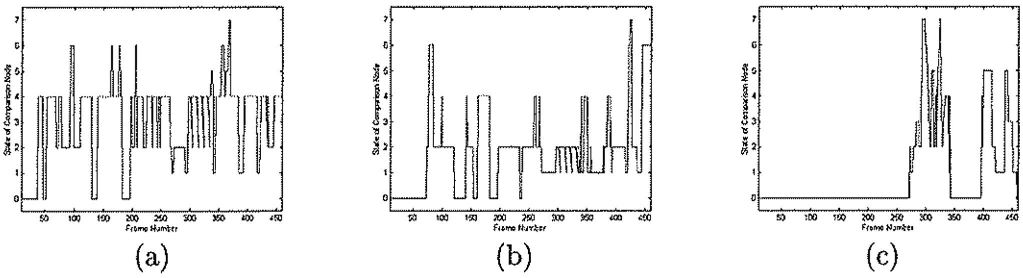
in the first comparison node, $S_A \leftarrow S_2$ in the second and $S_A \leftarrow S_3$ in the third. This situation can only occurs when the observed states (0-7) in 3 comparison nodes are the same and the inferred states (high and low) of the modality confidence nodes in all MUs (6 in total as landmark modality is implemented with 2 MUs) are the same. The probability of this situation occurring is $((\frac{1}{8})^3 \times (\frac{1}{2}))^6$ (i.e. 8.67e-19). Moreover, this situation can only happen when the prior over all states (A_1-A_6) in the correspondence node are the same. In the experiment, this situation was not seen.

Table 7.8 illustrates the CPT of the confidence indicator of the segmentation status of the highest point of the subjects. All other confidence indicator nodes have the same CPTs. Note that these tables (i.e. Tables 7.7 and 7.8) are manually generated based on the domain knowledge. The choice of the values are manually selected based on the performance when tested with different values. More appropriate values can be obtained based on estimation techniques (e.g. maximum likelihood estimation). The values used may be imperfect, but have acceptable results for the experiments conducted. They might need to be changed for other environments and different lighting conditions in order to make the results more reliable.

Observed states in the evidential nodes: Figures 7.21 and 7.22 show the observed states in the three comparison nodes and the three confidence indicator nodes respectively. These observations are obtained from the apparent height modality obtained from the two sequences taken by the two cameras in the tracking example (Figure 7.18).

Table 7.8: *Conditional probability table of a confidence indicator node given the states of its parent nodes. (M.C. = Modality Confidence).*

M. C.	low	medium	high
low	60%	30%	10%
high	10%	30%	60%



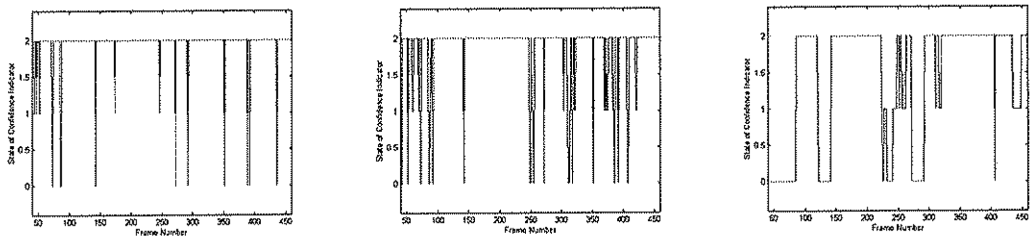
(a) (b) (c)

Figure 7.21: *The observed states in the comparison nodes.*

Figures 7.21(a-c) depict the observed states (see Table 7.4) in the three comparison nodes, comparing S_A, S_B, and S_C in the left camera image with all 3 subjects in the right camera image respectively. For example, Figure 7.21(a) depicts the observed states based on the comparison results of S_A with all 3 subjects. Note that the states remain 0 (as the subject in the left camera is not similar to any of the subjects in the right, see Table 7.4) in a comparison node until evidence is observed. For example, the state shown in Figure 7.21(c) remains 0 until the 269^{th} frame, when person 3 (i.e. S_C) first appears in the left camera image.

Figures 7.22(a-c) depict the states in the three confidence indicator nodes. These three indicators are (a) the segmentation status of the highest points (b) the segmentation status of the lowest points of the subjects, and (c) the mean difference of subjects' apparent heights. The states 0, 1 and 2 represent low, medium and high confidence respectively. From the whole sequence pair, it can be seen that the confidence level indicated by the highest point is in general higher than that of the lowest point. The less reliable segmentation of the lowest point may arise from the subject's shadow. This also suggests that the highest point can be more accurately extracted, and is thus better for use as a feature point to represent the image position of subjects for establishing correspondence.

Inferred probability distributions of the unobserved nodes: The observed visual evidence in the evidential nodes (i.e. comparison nodes and confidence indicator nodes)

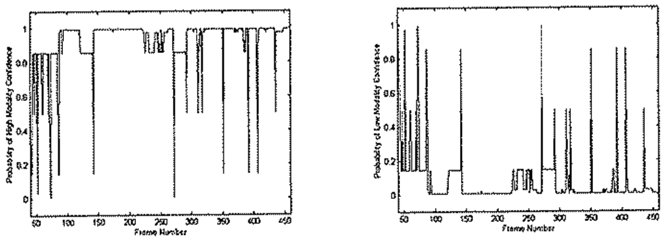


(a) (b) (c)

Figure 7.22: *The observed states in the confidence indicator nodes.*

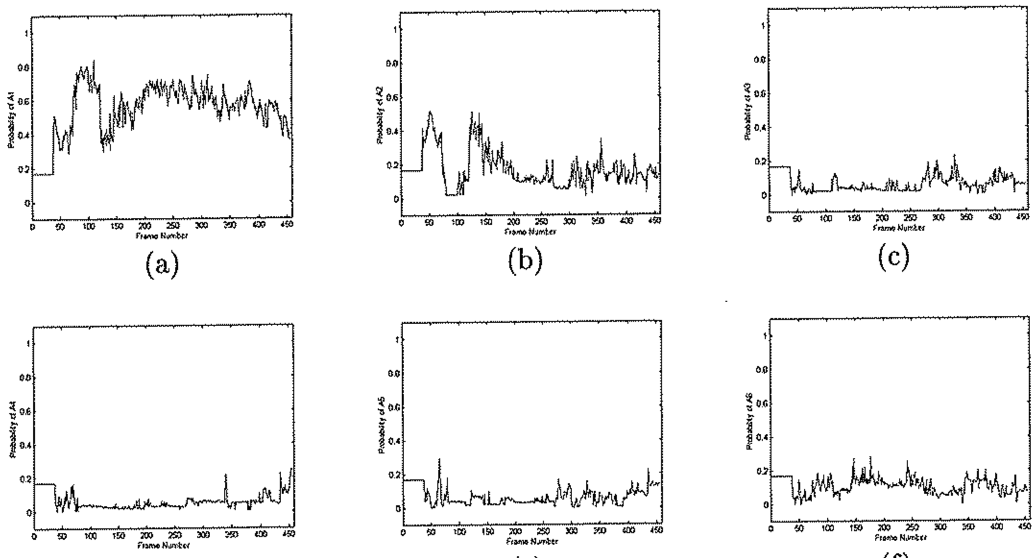
is used to infer the unobserved variables represented by the modality confidence node and the correspondence node. Figures 7.23 and 7.24 show the inferred probability distributions over the states of the modality confidence node and the correspondence node. Figures 7.23(a-b) show the inferred probability of the states (i.e. high confidence state in Figures 7.23(a) and low confidence state in Figures 7.23(b)) in the modality confidence node of the apparent height modality. Figures 7.24(a-f) show the inferred probability for the states of A_1-A_6 (see Table 7.3) in the correspondence node based on all modalities. Note that before the 38^{th} frame, the probabilities are 16.67% for all states, since there are no subjects visible in either camera image before this frame. Therefore, the system does not need to begin matching subjects between two camera images. The probabilities of all states were set equally (i.e. $\frac{1}{m!}$ as discussed in Section 4.4.1) by the system for this initial condition.

Obtaining the probability distribution over assignment combinations: Here, we explain how to obtain the results of subject correspondences when the numbers of subjects in both camera images are less than m (i.e. the maximum number of subjects in two camera images built in the BBN), and the number of subjects are different in the two images. For example, before the 120^{th} frame of the tracking example, there are two subjects, person 1 and person 2, in both camera images (see an example of 110^{th} frame in Figure 7.18(b)). After the 120^{th} frame, these two subjects meet in the left camera image, I_1, and are segmented as one single blob, but not in the right camera image, I_2 (see an example of the 130^{th} frame in Figure 7.18(c)). Since the number of subjects in I_1 has changed, the system interprets that there is matching ambiguity and the subject identity may be incorrect. From the MCCT, the system matches subjects across cameras in order to obtain the identity, i.e. "1" (of S_1) or "2"(of S_2), from I_2. Since the system does not rely on any occlusion reasoning techniques to resolve the ambiguity, it continues to match subjects across cameras (this ambiguity is resolved when the merged blob splits



(a) (b)

Figure 7.23: *The inferred probability distribution in the modality confidence node.*



(a) (b) (c)

(d) (e) (f)

Figure 7.24: *The inferred probability distribution in the correspondence node.*

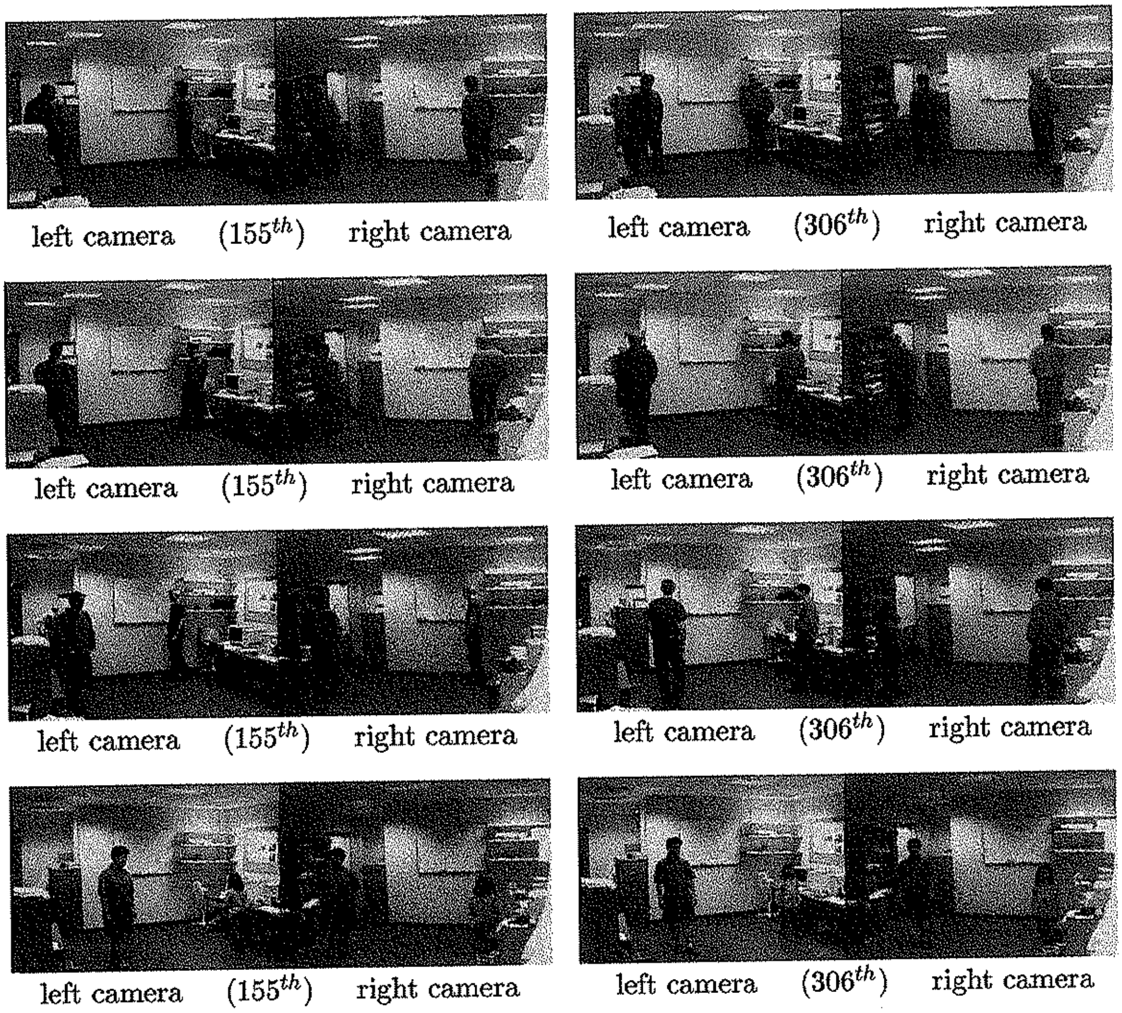
by passing of the identity from the camera where no ambiguity is present, and will be discussed later). The system needs to match the one subject (S_A, i.e. merger blob) in I_1 and two subjects (S_1 and S_2) in I_2 in order to determine whether S_A corresponds to S_1 or S_2, i.e. two matches: $S_A \leftarrow S_1$ and $S_A \leftarrow S_2$. The inferred probability distributions over the assignment combinations, A_1-A_6, at the 121^{st} frame are (0.519323 0.301932 0.022947 0.030193 0.102657 0.022947), as shown in Figure 7.24. The probability of the match $S_A \leftarrow S_1$ can be marginalised (see Equation (4.7)) from A_1 ($\{S_A \leftarrow S_1, S_B \leftarrow S_2, S_C \leftarrow S_3\}$) and A_2 ($\{S_A \leftarrow S_1, S_B \leftarrow S_3, S_C \leftarrow S_2\}$), while the match $S_A \leftarrow S_2$ can be marginalised from A_4 ($\{S_A \leftarrow S_2, S_B \leftarrow S_1, S_C \leftarrow S_3\}$) and A_5 ($\{S_A \leftarrow S_2, S_B \leftarrow S_3, S_C \leftarrow S_1\}$). Thus, the probability of $S_A \leftarrow S_1$ is 0.821255 (i.e. 0.519323 + 0.301932), and that of $S_A \leftarrow S_2$ is 0.13285 (i.e. 0.030193 + 0.102657). The system then assigns identity "1" (of S_1) to S_A (the merged blob), as can be seen in the left camera image in Figure 7.18(c).

After the 142^{nd} frame, the merged blob in I_1 splits into two blobs (see Figure 7.18(d)). The system assigns the label S_A to the left subject and S_B to the right subject in I_1. Since the number of subjects in I_1 has changed, the system matches subjects across cameras to obtain the identity (i.e. "1" or "2") from the subjects in I_2 for assigning to S_A and S_B, i.e. two assignment combinations : $\{S_A \leftarrow S_1, S_B \leftarrow S_2\}$ and $\{S_A \leftarrow S_2, S_B \leftarrow S_1\}$. The inferred probability distribution over the assignment combinations, A_1-A_6, are (0.474634 0.298221 0.033602 0.036870 0.021002 0.135670), as shown in Figure 7.24. The probability of the matches $\{S_A \leftarrow S_1, S_B \leftarrow S_2\}$ can be obtained from A_1 ($\{S_A \leftarrow S_1, S_B \leftarrow S_2, S_C \leftarrow S_3\}$), while that of the matches $\{S_A \leftarrow S_2, S_B \leftarrow S_1\}$ can be obtained from A_4 ($\{S_A \leftarrow S_2, S_B \leftarrow S_1, S_C \leftarrow S_3\}$). Thus the probability of $\{S_A \leftarrow S_1, S_B \leftarrow S_2\}$ is 0.474634, and that of $\{S_A \leftarrow S_2, S_B \leftarrow S_1\}$ is 0.036870. Therefore, the left subject, S_A, in I_1 is assigned the identity "1" (of S_1), and the right subject, S_B, is assigned the identity "2" (of S_2), as can be seen in the left camera image in Figure 7.18(d).

7.3.2 Performance Evaluation

To highlight the strength of Bayesian modality fusion for combining multiple cues, it is compared with a popular fusion method adopted by some tracking systems, e.g. [70, 23] as reviewed in Chapter 2. This method assumes all modalities are independent, and is often called the *naive Bayes* method. The matching result is based on similarity measurement computed by $M(\mathbf{S}, \mathbf{S}') = \prod_{k=1}^{n} P(a_k|a'_k)$, where $\mathbf{S}$ and $\mathbf{S}'$ represent two subjects to be matched with n different features a_k and a'_k respectively. In order to compare the robustness of these two methods, 20 sequence pairs (each sequence contains 500 frames) are collected with two people interacting with each other in the overlapping FOV. Figures 7.25, 7.26 and 7.27show some sample images of the sequence pairs that used for evaluating the matching results. Each row shows 2 images frames from a sequence pair.

The main aim of this experiment is to compare the performance of inter-camera subject correspondences based on two different methods (i.e BBNs and naive Bayes). When subjects are not visually isolated, there is no ground truth of the subject correspondence. For example, two subjects are imaged as one single blob in the left camera image (due to occlusion), and two blobs in the right. The single blob in the left camera might belong to either or both of the two subjects. Thus, no ground truth of subject correspondences exists when occlusion occurs. To evaluate the results of matching subjects in two cam-



left camera (155th) right camera

left camera (306th) right camera

left camera (155th) right camera

left camera (306th) right camera

left camera (155th) right camera

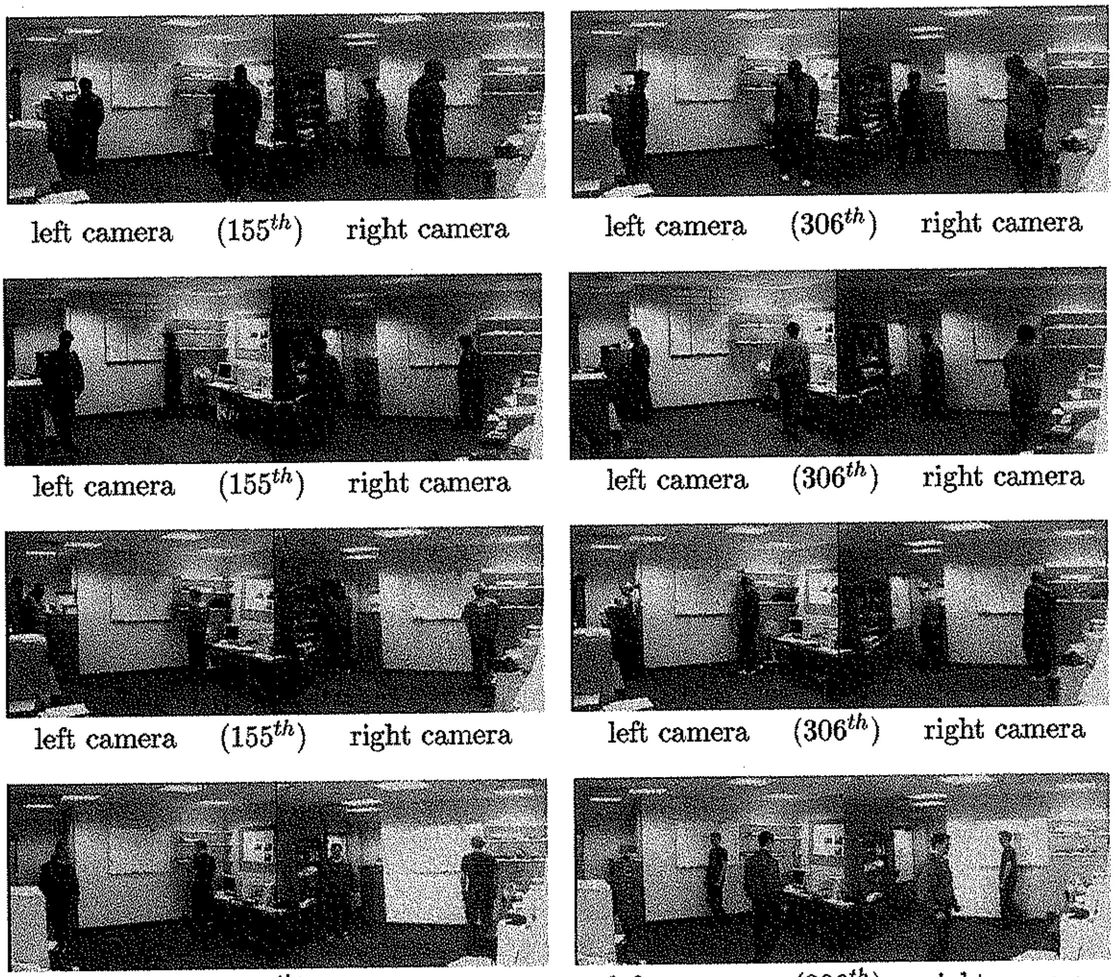
left camera (306th) right camera

left camera (155th) right camera

left camera (306th) right camera

Figure 7.25: *A tracking example.*

eras, when the subjects are visually isolated, the matching results are counted. Since the space of the room where the experiments were conducted is limited, only two people are used so as to limit the number of occluded frames. The tracking of two people is achieved by matching subjects across cameras in order to resolve the occlusion problem. The people (including man, women and children) had a wide range of heights, wore different coloured clothes and walked around the room randomly. Among people in these 20 sequences, the tallest is 193 cm and the shortest 124cm. The sequences were captured under a range of lighting conditions including day light with/without blinds, with/without artificial lights, and with/without daylight (i.e. at night). These 20 sequences were found cover acceptable different variations of subjects appearance and the lighting conditions in the room where the experiments were conducted. The comparison



left camera (155th) right camera

left camera (306th) right camera

left camera (155th) right camera

left camera (306th) right camera

left camera (155th) right camera

left camera (306th) right camera

left camera (155th) right camera

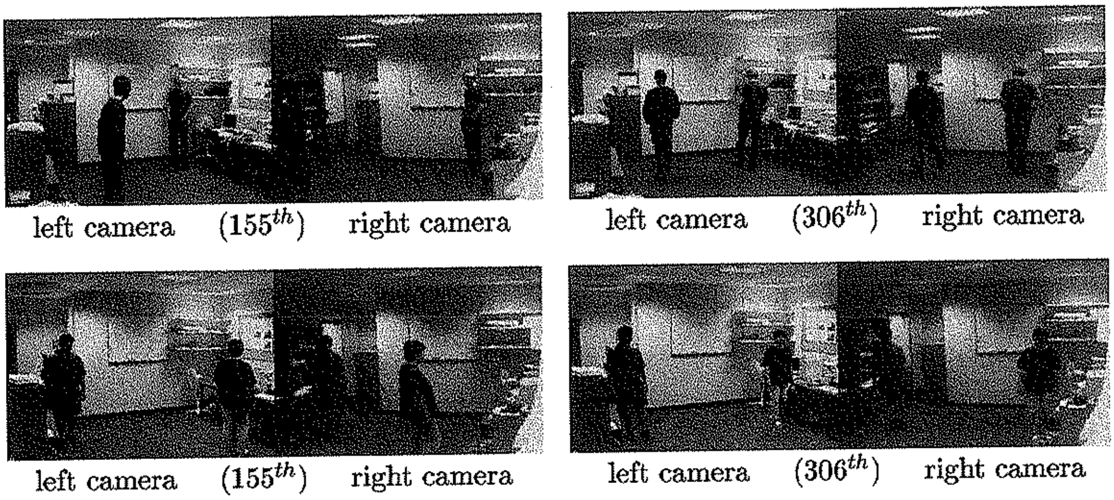
left camera (306th) right camera

Figure 7.26: *A tracking example.*

of these two methods is based on these 20 sequences.

To compare subjects' features based on the accumulated evidence (as discussed in Section 4.4.2, see Equation (4.15)), $\frac{1}{\sum_{i=0}^{q-1}\alpha_i}\sum_{i=0}^{q-1}\alpha_i\mathcal{F}(l-i)$ is used in the comparison nodes for all modalities, where q is set to 3, and α_0 is set to 3, α_1 to 2 for and α_2 to 1. Thus, the information from 3 consecutive frames are used to compare subjects' images.

Figure 7.28 illustrates the results of matching two people between two camera images. The accuracy rate of each sequence pair is the overall matching accuracy over all frames. Both methods use the matching results (i.e. a probability distribution over a combination of assignments) from previous frames as a prior in the current frame. Both methods also use the accumulated evidence (as discussed above) for comparison. The ground truth of matching (subject correspondences between two camera images) was generated by hand.



left camera (155th) right camera left camera (306th) right camera

left camera (155th) right camera left camera (306th) right camera

Figure 7.27: *A tracking example.*

When occlusion occurs in either or both camera images, the matching results are not counted. The average accuracy over all 20 sequence pairs is about 99.1% with standard deviation 1.2% for the Bayesian modality fusion method and 96.5 % with standard deviation 2.4% for the naive Bayes method. The accuracy range is between (96.3%, 100%) (i.e. (minimal, maximal)) for the Bayesian modality fusion method and (92.4%, 100%) for the naive Bayes method.

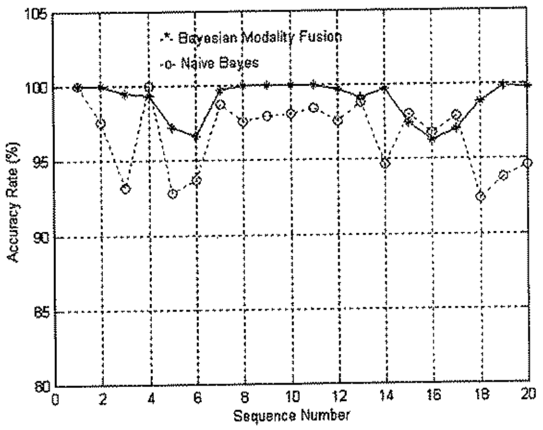


Figure 7.28: *The accuracy rate of matching subjects between two camera images based on Bayesian modality fusion and a naive Bayes method for 20 image sequence pairs.*

Incorrect matches can result from image noise, a less reliable estimated subject appearance (due to a different illumination condition from that used during mapping learning in training phase) and positional ambiguity of the extracted features. It was found that colour modality became less reliable in strong daylight. The reason for this is twofold. First, the learnt mapping used for estimation of subject appearances across cameras was obtained at night (without daylight). The estimate subject colour for inter-camera subject correspondences became less reliable under strong daylight. Second, the strong daylight can be reflected by clothes causing significant changes in colour distribution. This is because hue and saturation are in polar coordinates. When colour distribution change significantly can cause hue changes of up to 180 degrees in polar coordinates.

To further evaluate the improvement when using colour modality for inter-camera subject correspondences, the experiments were performed without using colour modality. Over 20 sequence pairs, the accuracy rate of matching subjects is about 95.4% with a standard deviation of 1.8% for the Bayesian modality fusion method when not using colour modality. Compared to the use of all modalities (the accuracy is about 99.1%), the difference (between 95.4% and 99.1%) indicates that the system achieves a 3.7% higher accuracy when using colour modality. The reason this improvement is not that high is probably because about half the people wore black and/or grey clothes. This causes the colour distribution unstable in the HS plane (such as hue changes up to 180 degree). This suggests that the use of other colour representations might result in a better accuracy.

The modalities of epipolar geometry is less reliable when the actual heights of people are similar. This is because epipolar lines are almost horizontal for the system camera setup. Therefore, the epipolar lines are close to each other when two people are of about the same heights and at the same depths. Since the distance between the computed epipolar line and highest point of the candidate matching subject is used as a match score, the system may not make a correct match due to the distance between the highest point of a subject and different epipolar lines are similar.

It was found that the line landmark modality was less reliable, when people are in the same VA in a camera image. This is because the rules (see Section 5.4.1) do not apply in this case. This suggests that the distance between two neighbouring line landmarks should be designed to be approximately equal to the width of a human body. Thus the system has more chance to view people in different VAs.

The results indicate that the BBN method is better in combining multiple visual modalities for matching subjects across cameras. This can be seen from the fact that the average accuracy of the BBN method is 2.6% higher than the naive Bayes method. Although the average accuracy is not much higher than the naive Bayes method, BBN method is also more stable than the naive Bayes method, as seen with the smaller standard deviation and smaller range.

To further evaluate the performance of these two methods, the third of the 20 sequence pairs, discussed above, is used. In this sequence pair, there are two people in both camera images. The goal is to match two subjects, S_A and S_B, in one camera image to S_1 and S_2 in the other image. The possible combination of assignments are $A_\alpha \in \{A_1, A_2\}$, where $A_1 = \{S_A \leftarrow S_1 \text{ and } S_B \leftarrow S_2\}$ and $A_2 = \{S_A \leftarrow S_2 \text{ and } S_B \leftarrow S_1\}$. The ground truth (i.e. subject correspondences between two camera images) is A_1 assignments. Figure 7.29 shows the inferred probability of A_1 based on BBN and naive Bayes. Since summation of probability of A_1 and A_2 is equal to one, the probability of A_1 being less than 0.5 represents an incorrect match. In this sequence pair, occlusion is present from the 155th to the 196th frame, where both A_1 and A_2 are set to 0.5.

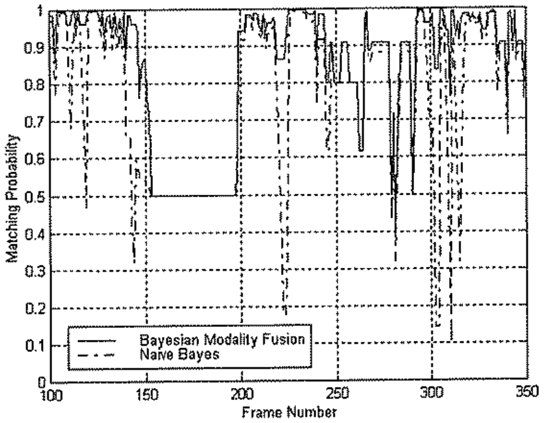


Figure 7.29: *The inferred probability of combination of assignments A_1 using Bayesian modality fusion and naive Bayes.*

Figure 7.30 shows the false rate (i.e. ratio of the number of frames of false matching to the total number of frames) at every frame instants across the whole sequence. Over the whole sequence pair, the false rate is less than 1% for BBN and about 8% for naive Bayes. To estimate the computational cost, the time consumption (in seconds) of both methods was recorded and is shown in Figure 7.31. Throughout the whole sequence, the computational time of the BBN method is 17% higher than that of the naive Bayes.

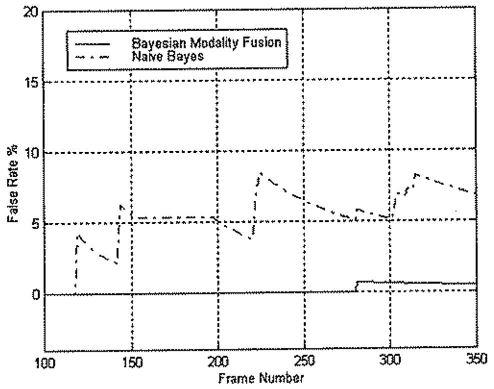


Figure 7.30: *The false rate of matching subjects between two camera images using Bayesian modality fusion and naive Bayes.*

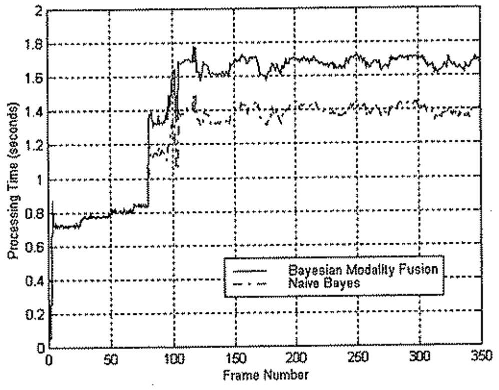


Figure 7.31: *The processing time for Bayesian modality fusion and naive Bayes.*

7.4 Summary

This chapter has discussed the experiments carried out regarding the matching task in MCCT. The first experiment involves selection of parameters of Gaussian variables for different modalities. Based on the obtained Gaussian parameter of apparent height modality, an example is also given to illustrate detection of matching ambiguity in MCCT. The second experiment focuses on the estimation of subject appearance across two camera images for recognition-based modalities. The third experiment investigates the application of Bayesain modality fusion for matching subjects across cameras. The discussion and conclusion of the work presented in this thesis will be given in the next chapter.

Chapter 8

Conclusion

8.1 Summary of Work

This thesis has developed a system for tracking multiple moving people in an indoor environment using two static, widely separated and un-calibrated cameras. The tracking system consists of two tracking modes: SCT (Single Camera Tracking) and MCCT (Multiple Camera Cooperative Tracking). In the MCCT mode, the system matches subjects' images across cameras to establish subject correspondences between the two camera images. Experimental results show that when performing MCCT, the system can track people with identities over time using two cameras cooperatively.

The SCT tracking mode was first discussed which includes two major steps: preprocessing and matching subjects between successive frames from a camera. Two stages of preprocessing are performed before the matching task: (1) segmentation of the moving subjects from the still background and (2) extraction of feature points from the segmented subjects' images. After preprocessing, the system establishes the feature correspondences between consecutive image frames using Kalman filters for tracking people.

In MCCT, a Bayesian belief network is used to adaptively fuse multiple modalities for matching subjects across cameras. Compared to SCT, the features used for subject correspondences in MCCT are extracted from different camera coordinate systems and from different physical processes, making matching more difficult. On the other hand, these different modalities regarding different constraints are highly correlated, since they are all related to the same scene. A framework, based on Bayesian modality fusion, is used to probabilistically infer subject correspondences between two camera images in order to handle data uncertainty and capture dependencies between different modalities.

Geometry-based modalities are used to handle significant image variations resulting from the features being obtained from two widely separated cameras. This problem is overcome by finding the geometric relationship between the highest points of corresponding subjects in two camera images. The homography induced by the virtual plane of a person is used to transfer a subject's highest point from one camera image to the other camera image in order to search for the corresponding subject. Epipolar geometry is used to constrain the highest point of corresponding subject across camera images on the epipolar line. Scene knowledge based on landmarks is also used for geometric reasoning about subject correspondences for MCCT.

To match subjects between camera images based on recognition-based modalities, the image patterns extracted from the subjects in one camera are used as a model for recognition of the corresponding subject in the other camera. The main difficulty of applying recognition-based modalities lies in wide variations in subject appearance due to changes in pose, scale and lighting condition, such that subjects appear different from different viewpoints. To compensate for the appearance variation, so as to make inter-camera subject correspondences more reliable, the mapping of subject appearance between two camera images is learnt using SVR (Support Vector Regression). By using the learnt mapping, one can estimate the appearance of the subject across camera images in order to obtain a "camera-dependent" model for matching subjects, though the mapping needs to be re-learnt if the camera is moved.

Finally, the accuracy of appearance estimation based on SVR, and the tracking system based on Bayesian modality fusion are both compared to different methods in order to demonstrate their robustness. The SVR method can estimate the subject appearance across cameras with a smaller error than the disparity in the appearance (i.e. without estimation) of corresponding subjects in two camera images. For 5 test sequence pairs, the error between estimated and observed apparent colour is 19.28% of the disparity of the apparent colour in two camera images, and for apparent height, the error between estimated and observed apparent height is 15.78% of the disparity of the apparent height in two camera images. Bayesian modality fusion achieved a 99.1% accuracy rate of matching subjects across camera images, 2.6% higher than the naive Bayes method (which assumes all modalities are independent). The tracking system presented in this thesis has been demonstrated to handle occlusion and maintain identities of multiple people consistently by using two cameras cooperatively.

8.2 Limitations

The methodology and different techniques for inter-camera subject correspondences proposed in this thesis are not free of caveats. Care must be taken in how they are applied. It is important for these limitations to be understood when building similar systems. The limitations discovered are listed below.

- In the geometry-based modalities, the highest point of a subject is used for representing the subject position in an image. Different modalities are then used to constraint the position of the corresponding points across camera images in order to match subjects between camera images. Similarly, in the recognition-based modalities, the highest point is used to determine the image position where the apparent colour of a subject's image is to be sampled; the highest and lowest points are used to determine the apparent height of a subject. Positional ambiguity in extracted features for representing a subject can degrade the matching reliability. This limitation is due to the inherent difficulty in feature extraction, as discussed in Chapter 1, that the features extracted from different camera images are in general corresponding to different parts of an object in the 3D world. To handle this issue, this thesis uses the segmentation status of the feature position between image frames in order to dynamically adjust the modality confidence. However, when the ambiguity in feature position exists for longer than two frames, the system may not detect this positional ambiguity. One possible remedy is the incorporation of a human shape model to segment the motion blob, in order to extract accurate feature points from a subject's image.

- The homography induced by the virtual plane of a person is used to transfer a subject's highest point from one camera image to the other camera image in order to search for the corresponding subject. However, homography only applies to the points lying on a scene plane. The homography modality may be less reliable for inter-camera subject correspondence when people change their poses significantly or when the ground is not a plane such that the subject's highest points do not lie on the same virtual plane. To improve subject correspondences based on this modality, the system can be made to recognise people's poses and hence make this method more reliable.

- The affine camera model is adopted to compute the epipolar geometry for searching for corresponding subjects between two camera images. This model only provides good approximation of the perspective model when the FOV is small and the variation in depth of the scene along the line of sight is small compared to its average distance from the camera. The assumption may not hold when people first enter the FOV. This is because the distance between the person and the camera is longer than the average distance from the camera when a person enters the room through the door . A similar situation could happen when a person enters the FOV from the corner where the camera is located. In this case, the distance between the person and the camera is shorter than the average distance from the camera.

- The landmark modality is used to geometrically reason the positions of corresponding subjects in two camera images. This constraint is a very computationally efficient algorithm. However, the people tracked in the environment must be viewed in front of the landmarks and in the overlapping area of FOVs. Moreover, the distance between two neighbouring line landmarks should be designed to be approximately equal to the width of a human body. Thus, the system has more chance to view people in different VAs in order to reason the position of corresponding subjects in two camera images. This is because more than one person could be in the same VA if the distance between two neighbouring line landmarks is too wide.

- A key issue in appearance estimation based on learnt mapping is that it may not handle different illumination conditions well. This could cause the estimation of subject appearances to be less reliable and result in incorrect matches. This issue is particularly serious in the modality of apparent colour, e.g. outdoor lighting can change colour significantly. As a result, the learnt mapping becomes less reliable. It might be useful to learn the correlation between the mappings for different illumination conditions in order to use different mappings for different illuminations.

- BBNs are well-established as representations of domains involving uncertain relations among several random variables. This thesis uses a BBN to fuse different visual modalities from two cameras for establishing inter-camera subject correspondences. The structure and the parameters (i.e. CPTs) of the network are designed from domain knowledge, and is only suitable for the camera setup used in this thesis. For different camera setups and different environments, different networks need to be designed.

- The goal of this thesis is to segment the images of moving subjects from the background and then to track moving (walking) people in the overlapping area over time from an image sequence pair of two monocular cameras. The system can handle the occlusion problem by using two widely separated cameras cooperatively. When occluded people are separated as two single blobs, the system passes subject identities across cameras by establishing inter-camera subject correspondences. Thus, the system can continuously track people by using two cameras cooperatively. The underlying assumption of the system is twofold. First, people are walking in the overlapping FOVs of two cameras. Second, the system always has one camera image where occlusion does not occur.

8.3 Future Work

There remain several avenues of interest to explore. The framework based on Bayesian modality fusion is quite general and as such it might be useful for different applications, such as tracking people walking in a large area monitored by multiple cameras (i.e. more than two cameras). The identities can then be maintained by matching subjects across cameras when people are in overlapping FOVs. The system could also be extended to recognise the activities of multiple people by taking advantage of consistent maintenance of identities. In order to coordinate all cameras to track multiple people, the communication protocols for multiple camera cooperative tracking require further study to make communication between cameras more reliable.

As pointed out in Chapter 2, the difficulty in tracking a group of people lies in the occlusion problem, where the system has no mechanism to update the visual information in order to track people. Individual camera can only cover a limited space and is subject to failure and/or measurement inconsistencies. The work presented in this thesis uses two cameras to track people and assumes that there is an unambiguous image of each individual in at least one camera at all times. However, an individual may be occluded in both cameras simultaneously. A larger scale cooperating multi-camera system potentially has more visual information than a two-camera system. Investigation into sensor fusion mechanisms to obtain global measurements for resolving the occlusion in all cameras might lead to some success and would be extremely useful.

Automatic tracking of people based on video cameras has now become important in many applications, especially in surveillance, e.g. detect police-designed "target faces", public areas and even inside the home to monitor domestic violence. Sophisticated software can automatically track people and detect some activities. To perform these visual tasks, the detection stage is critical in that incorrect detection can cause the system to fail to track people and recognise activities. For example, the occlusion problem can cause a visual surveillance system to have ambiguous information from all cameras. One possible method to interpret and understand a scene more robustly is to incorporate other types of sensors or mount cameras at different places in order to detect people without ambiguity. Once people are detected in different sensors independently or cooperatively, the system can then track them by fusing data from all sensors to make the tracking more robust.

Appendix A

Camera Models and Calibration

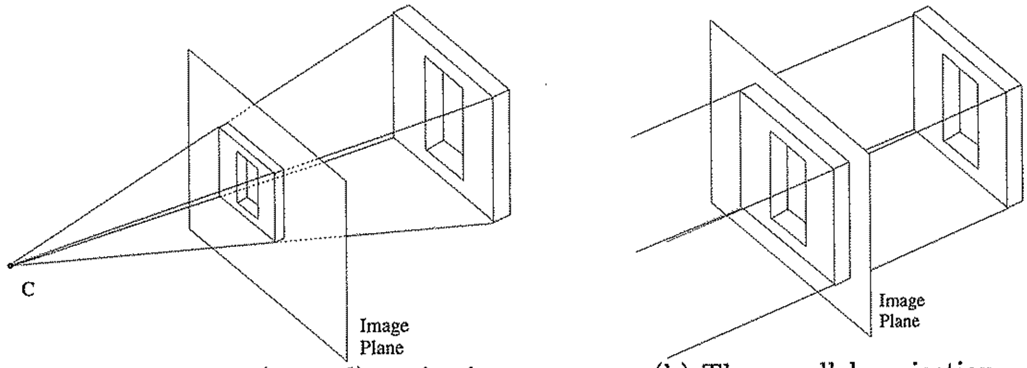
This appendix gives an introduction to the camera model and the concept of camera calibration. A camera is a mapping between the 3D world (scene space) and a 2D image. All cameras modelling central (perspective) projection (see Figure A.1), including the perspective camera and affine camera, are specialisations of the *projective camera* [59]. Firstly a description of the most general model, the projective camera, and camera calibration are given. Then, the perspective camera and the affine camera are described.

A.1 The Projective Camera and Camera Calibration

A camera projects a 3D world point $\mathbf{P} = (X, Y, Z)^{\mathrm{T}}$ onto a 2D image point $\mathbf{p} = (x, y)$. The mapping from $\mathcal{R}^3$ to $\mathcal{R}^2$ can be written in terms of a projective matrix $\mathbf{T} = [T_{ij}]$ in the homogeneous coordinates:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} & T_{13} & T_{14} \\ T_{21} & T_{22} & T_{23} & T_{24} \\ T_{31} & T_{32} & T_{33} & T_{34} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}, \tag{A.1}$$

where (x_1, x_2, x_3) and (X_1, X_2, X_3, X_4) are homogeneous coordinates related to $\mathbf{p}$ and $\mathbf{P}$, as $(x, y) = (x_1/x_3, x_2/x_3)$ and $(X, Y, Z) = (X_1/X_4, X_2/X_4, X_3/X_4)$. This camera model is termed a projective camera [109]. A projective camera places no constraint on the projection matrix $\mathbf{T}$ and the coordinated systems where $\mathbf{p}$ and $\mathbf{P}$ are measured, e.g. the world coordinate frame and the camera coordinate frame need not be orthogonal to the optical axis and these two frames need not be aligned [146]. This projective matrix can be computed directly from calibration [172] (i.e. computed from a set of world to image correspondences) and indirectly by computing a multiple view relation (e.g. fundamental matrix or trifocal tensor) [59]. This projective matrix $\mathbf{T}$ can be



(a) The perspective (central) projection.

(b) The parallel projection.

Figure A.1: *The camera projections.*

decomposed as follows [146, 48]:

$$\mathbf{T} = \mathbf{CEF} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ 0 & 0 & C_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} F_{11} & F_{12} & F_{13} & F_{14} \\ F_{21} & F_{22} & F_{23} & F_{24} \\ F_{31} & F_{32} & F_{33} & F_{34} \\ F_{41} & F_{42} & F_{43} & F_{44} \end{bmatrix}. \quad \text{(A.2)}$$

The 3×3 matrix $\mathbf{C}$ represents a 2D affine transformation (hence $C_{31} = C_{32} = 0$) and accounts for intrinsic camera parameters (i.e. geometric and optical characteristics). This matrix has a variable number of unknowns (usually up to 6) depending on the sophistication of the camera model. If there is no shearing (i.e. non-uniform scaling in some directions) in the camera axes and four parameters are used in $\mathbf{C}$:

$$\mathbf{C} = \begin{bmatrix} f\varepsilon & 0 & C_x \\ 0 & f & C_y \\ 0 & 0 & 1 \end{bmatrix}, \quad \text{(A.3)}$$

where f is the focal length (i.e. the distance between the image plane and the optical centre) ε is the camera aspect ratio (i.e. ratio of the horizontal and vertical pixel sizes) and (C_x, C_y) is the principal point (where the optical axis meets the image plane). These intrinsic parameters do not change as the position and orientation of the camera in space are changed. Knowledge of the intrinsic parameters allows one to perform metric measurements with a camera, i.e. to compute the angle between the rays determined by two pixels and the optical centre [48].

The 3×4 matrix $\mathbf{E}$ performs the projection operation. The 4×4 matrix $\mathbf{F}$ accounts for extrinsic camera parameters and encodes the relative position and orientation between the 3D world and 3D camera coordinate systems centred at the optical centre. In the linear projective equation (Equation (A.1)), the world point $\mathbf{P}$ (X_1, X_2, X_3, X_4) is first transformed to $\mathbf{P}_c$ in the 3D camera coordinate by $\mathbf{F}$ and then projected to the ideal (undistorted) 2D image coordinate by $\mathbf{E}$. Finally, it is transformed to the real image point (x_1, x_2, x_3) by the matrix $\mathbf{C}$.

To calibrate a single camera, one needs to determine the intrinsic and/or extrinsic parameters. From calibration, one can infer 3D information from the image coordinates (or *vice versa*) [172]. Calibration of multiple cameras requires the calibration objects for each camera to be measured in the same 3D world coordinate system in order to determine the relative positions and orientation [114].

A.2 The Perspective Camera

The perspective camera model is a specialisation of the projective camera. Figure A.1(a) shows the perspective projection onto an image plane where all projection rays converge at the camera centre C. This camera models the ideal perspective projection (i.e. no distortion) and is the familiar *pinhole camera* in which the 3D camera and 3D world coordinate frames are related by a rigid transformation [146]:

$$\mathbf{P}_c = \mathbf{RP} + \mathbf{t}, \tag{A.4}$$

where $\mathbf{P}$ is a point in the 3D world coordinate, $\mathbf{P}_c$ is a point in the 3D camera coordinate, $\mathbf{R}$ is a 3×3 rotation matrix (with rows $\{\mathbf{R}^{1\mathrm{T}}, \mathbf{R}^{2\mathrm{T}}, \mathbf{R}^{3\mathrm{T}}\}$) representing the orientation of the camera coordinate frame, and $\mathbf{t} = (t_x, t_y, t_z)^{\mathrm{T}}$ is a 3×1 translation vector representing the origin of the world coordinate frame. As a consequence,

$$\mathbf{F} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}. \tag{A.5}$$

Using Equation (A.3) the projective matrix $\mathbb{T}$ of the perspective camera can be written as:

$$\mathbf{T}_p = \begin{bmatrix} f\varepsilon\mathbf{R}^{1\mathrm{T}} + C_x\mathbf{R}^{3\mathrm{T}} & f\varepsilon t_x + C_x t_z \\ f\mathbf{R}^{2\mathrm{T}} + C_y\mathbf{R}^{3\mathrm{T}} & f t_y + C_y t_z \\ \mathbf{R}^{3\mathrm{T}} & t_z \end{bmatrix}. \tag{A.6}$$

A.3 The Affine Camera

An affine camera has a camera projective matrix $\mathbf{T}$ in which the last row has the form $(0, 0, 0, T_{34})$, i.e.

$$\mathbf{T}_{aff} = \begin{bmatrix} T_{11} & T_{12} & T_{13} & T_{14} \\ T_{21} & T_{22} & T_{23} & T_{24} \\ 0 & 0 & 0 & T_{34} \end{bmatrix}. \tag{A.7}$$

This affine camera model corresponds to a projective camera with its optical centre on the plane at infinity. As a consequence, all projection rays are parallel. Figure A.1(b) shows this parallel projection onto an image plane where all projection rays are parallel. In terms of image and world coordinates, the affine camera can be written as:

$$\mathbf{p} = \mathbf{M}\mathbf{P} + \mathbf{d}, \tag{A.8}$$

where $\mathbf{M}$ is a 2×3 matrix with elements $M_{ij} = T_{ij}/T_{34}$ and $\mathbf{d} = (T_{14}/T_{34}, T_{24}/T_{34})^{\mathrm{T}}$ is a 2D vector. The affine projection preserves the parallelism, i.e. lines that are parallel in the world remain parallel in the image.

Appendix B

Implementing Kalman filters

The parameters given in this appendix are used in this thesis and are adopted from McKenna et al. [103] which provide satisfactory results given in Section 3.3. The filter can follow the highest point well in the scenario when people walk in the office. The parameters selected in the model might need to be changed for different environments and people with different motions (e.g. running). The state vectors of the system represent the position, velocity and acceleration, in x and y coordinates, of the target's highest point respectively:

$$\mathbf{s}_x(k) = [x, \dot{x}, \ddot{x}]^{\mathrm{T}}, \tag{B.1}$$

$$\mathbf{s}_y(k) = [y, \dot{y}, \ddot{y}]^{\mathrm{T}}. \tag{B.2}$$

where $\dot{x}$ is defined as $\delta x/\delta t$, $\ddot{x}$ is defined as $\delta\dot{x}/\delta t$ and δt is the time step which is set to 1 between two consecutive frames. The δx is the difference between a tracked subject's highest points in two consecutive frames and the $\delta\dot{x}$ is the difference between the velocities of a tracked subject's highest point in two consecutive frames.

For both coordinates, the system dynamic of the target uses a second order model with constant acceleration [168]:

$$\mathbf{F}(k) = \begin{pmatrix} 1 & \delta t & \delta t^2/2 \\ 0 & 1 & \delta t \\ 0 & 0 & 1 \end{pmatrix}. \tag{B.3}$$

The system noise model $\mathbf{G}(k)$ for both x and y coordinates is defined as:

$$\mathbf{G}(k) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{B.4}$$

The system noise covariance $\mathbf{Q}(k)$ for both x and y coordinates is set as:

$$\mathbf{Q}(k) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}. \tag{B.5}$$

The system states $\mathbf{s}_x(k)$ and $\mathbf{s}_y(k)$ are defined as the same as the system measurement $\mathbf{z}_x(k)$ and $\mathbf{z}_y(k)$. Thus, the measurement model $\mathbf{H}$ for both x and y coordinates is defined as:

$$\mathbf{H}(k) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{B.6}$$

and the measurement noise covariance $\mathbf{R}(k)$ for both x and y coordinates is set as:

$$\mathbf{R}(k) = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.2 \end{pmatrix}. \tag{B.7}$$

The initial state of the x coordinate is set as the initial measurement:

$$\mathbf{s}_x(0|0) = [x, \dot{x}, \ddot{x}]^{\mathrm{T}}, \tag{B.8}$$

and for the y coordinate it is set as:

$$\mathbf{s}_y(0|0) = [y, \dot{y}, \ddot{y}]^{\mathrm{T}}. \tag{B.9}$$

The initial state covariances $P(0|0)$ are set as:

$$\mathbf{P}_x(0|0) = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1.5 \end{pmatrix}, \tag{B.10}$$

$$\mathbf{P}_y(0|0) = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 2.3 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{B.11}$$

Appendix C

Building the Secondary Structure

This appendix provides a method to build the secondary structure, i.e. junction tree, from the DAG (Directed Acyclic Graph) of the BBN (Bayesain Belief Network). This graph transformation includes building a number of intermediate structures which are known as a *moral graph* and a *triangulated graph*. The next step is to build the junction tree from the triangulated graph by identifying subsets of nodes (cliques) in the triangulated graph and inserting separators [76, 66].

C.1 Junction Tree Properties

The graphical and numerical properties of the secondary structure [78] of a BBN, defined over a set of n variables $\mathbf{V} = \{V_1, \dots, V_n\}$, are described as follows.

- The clusters satisfy the junction tree properties, which are
  1. given two clusters $\mathbf{P}$ and $\mathbf{Q}$ in $\mathcal{T}$ (i.e. the secondary structure of a BBN) all clusters on the path between $\mathbf{P}$ and $\mathbf{Q}$ contain the variables of $\mathbf{P} \cap \mathbf{Q}$, and
  2. for each variable $V_i \in \mathbf{V}$, the *family* of V_i (i.e. $\mathbf{F}_{V_i} = V_i \cup \Pi_{V_i}$) is included in at least one of the clusters, where Π_{V_i} are the parent node(s) of V_i.
- Each cluster $\mathbf{C}$ (and each separator $\mathbf{S}$) is associated with a potential $\phi_{\mathbf{C}}$ ($\phi_{\mathbf{S}}$) that maps each instantiation of $\mathbf{c}$ ($\mathbf{s}$) to a real number.
- For each cluster $\mathbf{C}$ and neighbouring separator $\mathbf{S}$, it holds that:

$$\sum_{\mathbf{C} \setminus \mathbf{S}} \phi_{\mathbf{C}} = \phi_{\mathbf{S}}. \tag{C.1}$$

When a cluster C and a neighbouring separator S satisfy Equation (C.1), $\phi_{\mathbf{S}}$ is said to be *consistent* with $\phi_{\mathbf{C}}$. When all pairs of clusters and neighbouring separators are consistent, the secondary structure is said to be *locally consistent.*

C.2 The Moral Graph

In building the secondary structure of the BBN, the network is first moralised. Given a DAG of a BBN, $\mathcal{G}$, the moral graph, $\mathcal{G}_m$, that corresponds to $\mathcal{G}$ is constructed by:

1. constructing an undirected graph $\mathcal{G}_u$ by dropping the directions of the edges in $\mathcal{G}$, and
2. constructing a moral graph $\mathcal{G}_m$ by adding undirected edges between each pair of nodes in Π_{V_i} (i.e. parent nodes of V in $\mathcal{G}$) for each variable V.

The right diagram in Figure C.1 shows the moral graph $\mathcal{G}_m$ of the DAG $\mathcal{G}$ in Figure 4.5 (shown in Section 4.3.3 and also plotted on the left side of the Figure C.1 for convenience) constructed by adding the undirected edges in $\mathcal{G}$. Added undirected edges are shown as dashed lines.

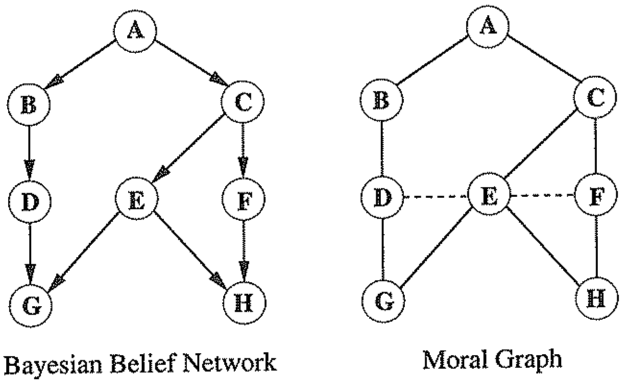


Figure C.1: *The moral graph constructed from the Bayesain belief network.*

C.3 The Triangulated Graph

After building the moral graph, the next step is to construct the triangulated graph. An undirected graph triangulated is a triangulated graph, $\mathcal{G}_t$, if and only if every cycle of length (i.e. number of edges in a cycle) >3 contains an edge that connects two non-adjacent nodes in the cycle. Figure C.2 shows the triangulated graph $\mathcal{G}_t$ of DAG $\mathcal{G}$, constructed by adding the edges in $\mathcal{G}_m$ of Figure C.1. The triangulated graph $\mathcal{G}_t$ that corresponds to $\mathcal{G}$ is constructed from the following rules.

1. Make a copy of $\mathcal{G}_m$ as $\mathcal{G}'_m$.
2. Perform the following steps repeatedly until there is no node left in $\mathcal{G}'_m$ (see Table C.1):
    - Select a node V in $\mathcal{G}'_m$.
    - Form a cluster composed of V and its neighbouring nodes in $\mathcal{G}'_m$. Add an edge between any pair of nodes in this cluster if this edge is not in $\mathcal{G}'_m$.
    - Add an edge in $\mathcal{G}_m$ corresponding to the new edge in $\mathcal{G}'_m$ added in last step.
    - Remove the node V from $\mathcal{G}'_m$.

The resulting $\mathcal{G}_m$ in now a triangulated graph $\mathcal{G}_t$. The dashed lines in Figure C.2 indicate the edges added to triangulate the moral graph. In general, there is more than one way to obtain the triangulated graph $\mathcal{G}_t$ of a DAG. The node-elimination carried out to remove the variables in the moral graph, according to step 2 above, is shown in Table C.1. The added edges and the induced clusters are also given.

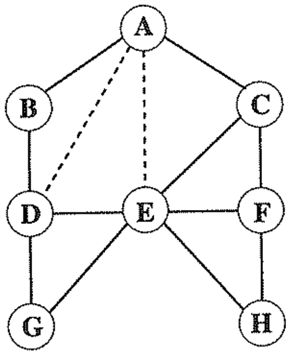


Figure C.2: *The triangulated graph constructed from the moral graph in Figure C.1.*

C.4 Identify the Cliques

After constructing the triangulated graph, the next step is to identify the cliques in the triangulated graph [91]. A clique is a subset of $\mathbf{V}$ in the DAG $\mathcal{G}$. Cliques can be extracted from the triangulation process by saving each induced cluster that is not a subset of any previously saved cluster. The cliques in $\mathcal{G}$ can be identified from Figure C.2 and Table C.1 as $\{DEG, EFH, CEF, ACE, ABD, ADE\}$.

Table C.1: *Elimination steps in triangulation of the moral graph in Figure C.1.*

Eliminated variables	Added edges	Induced clusters
G	none	DEG
H	none	EFH
F	none	CEF
C	$\overline{AE}$	ACE
B	$\overline{AD}$	ABD
D	none	ADE
E	none	AE
A	none	A

C.5 The Secondary Structure

The first step in building the secondary structure of the DAG, $\mathcal{G}$, is to set the cliques as clusters. The clusters are then connected to form an undirected tree and the appropriate separators are inserted. Separators are intersections of adjacent clusters, i.e. $\mathbf{C}_i \cap \mathbf{C}_j$. The secondary structure of the BBN is shown in Figure C.3.

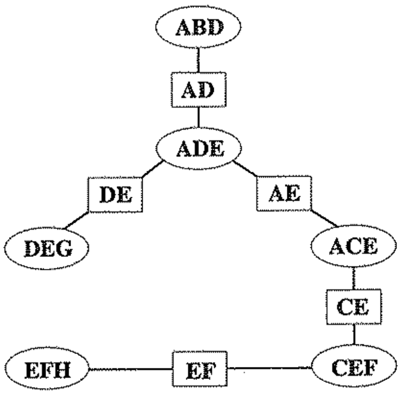


Figure C.3: *The secondary structure of the Bayesian belief network on the left side of Figure C.1.*

Note that in general there are several ways to triangulate the moralised graph. Finding the triangulation with the smallest number of cliques, for saving the representation and computation, is NP-hard [77]. However, the graph transformation process only needs to be performed once off-line. Jensen and Jensen [77] have shown that any exact inference algorithm based on local computations is at least as hard as the junction tree algorithm, and thus also NP-hard.

Appendix D

Inference in Bayesian Belief Networks

This appendix provides the inference procedures for computing $p(V|\mathbf{e})$ for sets of evidences $\mathbf{e}$ in the secondary structure of a BBN [76, 66]. The BBN is defined over a set of variables $\{V_1, \cdots, V_n\}$. The related definitions of notation and algebra are given in Section 4.3.4. For convenience, Figure 4.8 is shown in Figure D.1 again, where the dotted path indicates the control of the inference procedures with the dynamic observations.

D.1 Initialisation

After transforming the DAG of a BBN to a secondary structure (see Appendix C), the next step is to quantify the junction tree with potentials $\phi_{\mathbf{C}}$ and $\phi_{\mathbf{S}}$. The following procedure assigns the initial potential of each cluster and separator in the secondary structure using the given CPTs (Conditional Probability Tables) which are the $P(V_i|\Pi_{V_i})$ defined in a BBN.

1. Set each element $\phi_{\mathbf{C}}(\mathbf{c})$ in the potential of each cluster, and each element $\phi_{\mathbf{S}}(\mathbf{s})$ in the potential of each separator to 1:

$$\phi_{\mathbf{C}}(\mathbf{c}) \leftarrow 1, \tag{D.1}$$

$$\phi_{\mathbf{S}}(\mathbf{s}) \leftarrow 1. \tag{D.2}$$

2. Assign to each variable V a cluster $\mathbf{C}$ which contains the family of V (i.e. $V \cup \Pi_V$). This cluster is called the *parent cluster* of $\mathbf{F}_V$. Then multiply $\phi_{\mathbf{C}}$ by $P(V|\Pi_V)$:

$$\phi_{\mathbf{C}}(\mathbf{c}) \leftarrow \phi_{\mathbf{C}}(\mathbf{c})P(V|\Pi_V). \tag{D.3}$$

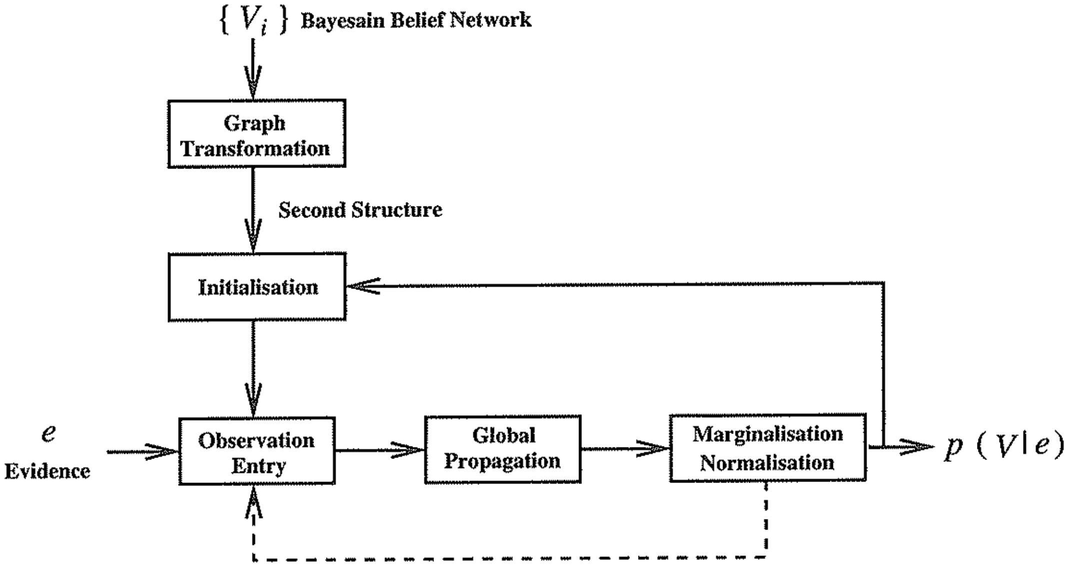


Figure D.1: *Block diagram of probabilistic inference in the secondary structure.*

3. Assign to each variable V a *likelihood*, denoted as ψ_V, which is a potential of $\{V\}$ and is used for entering the observations. This likelihood ψ_V maps each value v to a real number. Set each likelihood element $\psi_V(v)$ to 1:

$$\psi_V(v) \leftarrow 1. \tag{D.4}$$

After initialisation, the $P(V|\Pi_V)$ of each variable V has been multiplied into the potential of a cluster, and all separator potentials remain as $\phi_S \leftarrow 1$, such that the probability distribution represented by the tree is:

$$P(\mathbf{V}) = \frac{\prod_i \phi_{\mathbf{C}_i}}{\prod_j \phi_{\mathbf{S}_j}} = \frac{\prod_k P(V_k|\Pi_{V_k})}{1}. \tag{D.5}$$

From this equation, it can be seen that after initialisation the joint distribution (Equation (4.9)) represented by the secondary structure is the same as that in Equation (4.3) represented by the BBN. Since after initialisation of potentials the structure does not meet Equation (C.1), the result is a locally inconsistent structure.

D.2 Observation Entry

When collections of evidences $\mathbf{e}$ are received from some evidential variables $\{V\}$ with evidence values $\{v\}$, the incorporation of each evidence $V = v$ is achieved by encoding the evidence as a likelihood and entering this likelihood into the tree. This is explained as follows:

1. Incorporate the evidence $V = v$ (if $V \in \mathbf{E}$) as likelihood ψ_V^*:

$$\psi_V^*(v) = \begin{cases} 1 & , \text{ when } v \text{ is the observed state of } V. \\ 0 & , \text{ otherwise.} \end{cases} \tag{D.6}$$

2. Update the potential of a cluster that contains V:

$$\phi_{\mathbf{C}} \leftarrow \phi_{\mathbf{C}}\psi_V^*. \tag{D.7}$$

After entering the evidence, $\mathbf{e}$, for those clusters with evidences, the potentials $\phi_{\mathbf{C}}$ (which represent $P(\mathbf{C})$ as Equation (4.10)) have been modified to contain the evidences representing $P(\mathbf{C}, \mathbf{e})$.

D.3 Global Propagation

Having entered the observations, the next step is to perform global propagation in order to make the structure locally consistent. Global propagation consists of a series of local computations, called *message passes*, on the tree potentials that occur between two neighbouring clusters. A message passing from $\mathbf{C}_i$ to $\mathbf{C}_j$ forces the potential of the intervening separator, $\phi_{\mathbf{S}}$, to be consistent with $\phi_{\mathbf{C}_i}$ (see Equation (C.1)). Global propagation causes each cluster to pass a message to each of its neighbours and makes each cluster-separator pair consistent. Thus, the tree is locally consistent. In the following, firstly a description of a single message pass between two neighbouring clusters is given before multiple messages in the tree are considered.

D.3.1 Single Message Pass

Given two clusters $\mathbf{C}_i$ and $\mathbf{C}_j$ together with their intervening separator $\mathbf{S}$, a single message pass from $\mathbf{C}_i$ to $\mathbf{C}_j$ is achieved by performing:

1. *Message projection*: Save the old potential as $\phi_{\mathbf{S}}^*$ and assign a new potential $\phi_{\mathbf{S}}$:

$$\phi_{\mathbf{S}}^* \leftarrow \phi_{\mathbf{S}}, \tag{D.8}$$

$$\phi_{\mathbf{S}} \leftarrow \sum_{\mathbf{C}_i \backslash \mathbf{S}} \phi_{\mathbf{C}_i}. \tag{D.9}$$

2. *Message absorption*: Assign a new potential to cluster $\mathbf{C}_j$ using both the new and old potentials of separator $\mathbf{S}$:

$$\phi_{\mathbf{C}_j} \leftarrow \phi_{\mathbf{C}_j} \frac{\phi_{\mathbf{S}}}{\phi^*_{\mathbf{S}}}. \tag{D.10}$$

D.3.2 Multiple Message Pass

Given a junction tree, global propagation begins by arbitrarily choosing a cluster $\mathbf{C}_i$ and then performing message passes, including the use of two algorithms, *CollectEvidence* and *DistributeEvidence*. In the CollectEvidence phase, each cluster passes messages to its neighbouring clusters in $\mathbf{C}_i$'s direction. These passes begin from the cluster farthest from $\mathbf{C}_i$. In the DistributeEvidence phase, each cluster passes messages to its neighbouring clusters starting from $\mathbf{C}_i$ and moving away.

In the global propagation, each cluster passes its information to all other clusters in the tree. Thus, the encoded evidences in the potentials of some clusters (which include evidential variables) are passed throughout the tree.

After global propagation, the potentials of all clusters and separators will have been modified as $P(\mathbf{C}, \mathbf{e})$ and $P(\mathbf{S}, \mathbf{e})$. Then from the marginalisation (Equation (4.13)) and normalisation (Equation (4.14)), one can obtain $p(V|\mathbf{e})$, as described in Section 4.3.4.

Appendix E

Hue, Saturation and Value Colour Model

This appendix provides an introduction to the HSV colour model. There are many colour spaces which can be used to describe the colours, such as the red, green, and blue (RGB), hue, saturation and value (HSV), cyan, yellow and megenta (CYM), and hue, lightness and saturation (HLS).

The best known model is the RGB. It is used by most image acquisition hardware, with R, G and B being real numbers from the interval [0,1], representing the red, green and blue components, respectively. In order to represent the colour of an object with less sensitivity to intensity (brightness), it is advantageous that the representation of colour tone is separated from the intensity. The HSV model proposed in [152] meets this requirement, since the value defines the intensity. The hue is associated with the dominant wavelength in a mixture of light waves and defines the object colour (e.g. red or blue). Saturation describes the purity of the colour; the more the light reflected from an object is diluted by white light, the lower the saturation. For example, pink (red and white) is less saturated than pure red. A description of transformations between RGB and HSV colour space can be found in [74].

Figure E.1 shows the HSV colour space. The graph on the right side is the HS space obtained by dropping the Value component. The vertices of the hexagon at the top level represent red, yellow, green, cyan, blue and magenta. The root of the hexcone is defined by v=0 (corresponding to black) at the bottom and v=1 (corresponding to white) at the top level. The hue is measured by the angle around the root vertical axis. The saturation is measured as the ratio ranging from 0 on the root axis to 1 on the triangular sides of

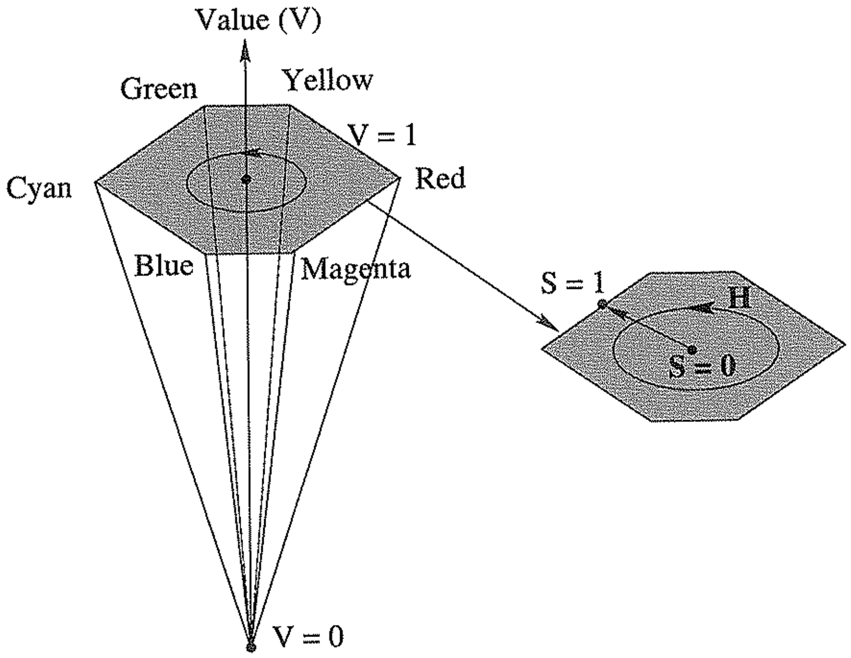


Figure E.1: *Hue Saturation and Value (HSV) colour space.*

the hexcone. By dropping the intensity component (i.e. V) to obtain a limited level of intensity invariance, the HS space is equivalent to a level of the hexagon.

Appendix F

Support Vector Regression

Support Vector Regression (SVR) was recently developed by Vapnik and co-workers [176, 44, 154, 153]. This statistical learning algorithm has been of great interest in the research areas of machine learning and pattern recognition, and has found many applications, such as head pose estimation [95] and signal detection [137].

The goal of the SVR algorithm is to achieve the nonlinear regression estimate in the input space by constructing a linear regression function in a high dimension feature space, where the input pattern $\mathbf{x}$ is mapped to the feature space via Φ. Thus, given a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, with input patterns $\mathbf{x}_i \in \mathcal{R}^n$ and interpretation $y_i \in \mathcal{R}$, the SVR problem can be defined as the determination of a function $f(\mathbf{x})$ which approximates an unknown desired function. It has the following form [176]:

$$f(\mathbf{x}) = \mathbf{w} \cdot \Phi(\mathbf{x}) + b, \tag{F.1}$$

where $\mathbf{w} \in \mathcal{R}^n$ is an unknown, "$\cdot$" denotes the dot product, and b is the unknown threshold. If the interpretation y only takes values -1 and $+1$, the learning problem is referred to as *support vector classification.* Otherwise, if the domain of y includes continuous real values, it is SVR.

By introducing a kernel function

$$K(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}), \tag{F.2}$$

the SVR problem can then be formulated as maximising the quadratic form defined as [44, 154]:

$$\begin{aligned} W(\alpha^*,\alpha) &= -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(\mathbf{x_i},\mathbf{x_j}) \\ &\quad -\varepsilon\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)+\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i), \qquad \text{(F.3)} \\ \text{subject to} &\quad \sum_{i=1}^{l}(\alpha_i^*-\alpha_i)=0, \qquad \text{(F.4)} \\ &\quad 0\le\alpha_i^*,\alpha_i\le C, \qquad \text{(F.5)} \end{aligned}$$

where α_i and α_i^* represent the parameters of the learning machine and C is a regularisation penalty factor to control the trade-off between the model complexity and the accuracy of the function. By maximising $W(\alpha^*,\alpha)$, one can obtain the coefficients $\{\alpha_i$ $\alpha_i^*\}$.

Additionally, from the derivatives of the Lagrange function [154], one can obtain $\mathbf{w}=\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)\Phi(\mathbf{x}_i)$. Then, by substituting this result into Equation (F.1) and from Equation (F.2) we can obtain:

$$f(\mathbf{x})=\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)K(\mathbf{x},\mathbf{x}_i)+b. \qquad \text{(F.6)}$$

where b is a threshold.

It is interesting to note that only a few parameters, α and α^*, take non-zero values, i.e. only those "important" examples, known as *Support Vectors* (SVs), are selected to construct the optimal approximation function (F.6), which is a linear combination of the SVs in high-dimensional feature space. However, instead of computing the map Φ explicitly, one only needs to compute the kernel function (F.2), done with greater ease.

In the experiment, the quadratic optimisation problem is solved by a decomposition algorithm based on the LOQO algorithm [175]. A Gaussian kernel

$$k(x_1,x_2)=exp(-\frac{\|x_1-x_2\|^2}{2\delta^2}) \qquad \text{(F.7)}$$

is used to build the SVR. The tolerance coefficient ε is used to define the ε-insensitive loss function [176] in SVR problems, such that the regression function has at most ε deviation from the actual interpretation:

$$|f(\mathbf{x})-y|=\begin{cases} 0, & \text{if } |f(\mathbf{x})-y|\le\varepsilon \\ |f(\mathbf{x})-y|, & \text{otherwise} \end{cases} \qquad \text{(F.8)}$$

where f is the regression function, and y is the interpretation of input pattern $\mathbf{x}$. Normally, ε can be used to control the accuracy of a SVM regressor. A large value of ε may lead to a regression function with poor accuracy and good real-time performance since a large error is acceptable by the loss function (Equation (F.8)) and a small number of SVs can be obtained from training. However, a small value of ε can result in overfitting to the training set. The related parameters selected in the experiment are given in Section 7.2.

Bibliography

[1] M. Abdulghafour and M. A. Abidi. Data fusion through non-deterministic approaches - a comparison. In *Proc. of the SPIE: Sensor Fusion*, volume 2059, pages 37–53, 1993.

[2] J. K. Aggarwal. *Multisensor Fusion for Computer Vision.* Springer-Verlag, 1993.

[3] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999.

[4] J. K. Aggarwal, L. S. Davis, and W. N. Martin. Correspondence processes in dynamic scene analysis. *Proc. of the IEEE*, 69(5):562–572, 1981.

[5] R. Alferez and Y. F. Wang. Geometric and illumination invariants for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:505–536, 1999.

[6] C. S. Andersen. An analysis of five depth recovery techniques. In *Proc. of the First Nordic Summer School on Active Vision and Geometric Modelling*, pages 93–104, Aalborg, Denmark, 1992.

[7] M. Armstrong. *Self-Calibration from Image Sequences.* PhD thesis, University of Oxford, 1996.

[8] E. P. Baenen. Generalized probabilistic reasoning and empirical studies on computational efficiency and scalability. Master's thesis, Graduate School of Engineering, Air Force Institute of Technology, Ohio, USA, December 1994.

[9] Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association.* Academic Press Limited, San Diego, CA, 1988.

[10] K. Barnard. Improvements to gamut mapping colour constancy algorithms. In *European Conference on Computer Vision*, pages 390–403, 2000.

[11] M. Berger and G. Danuser. Deformable multi template matching with application to portal images. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 374–379, June 1997.

[12] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. In *Proc. of the 12th Conference on Uncertainty in AI*, pages 115–123, Portland, Oregon, USA, August 1996.

[13] K. Brammer and G. Siffling. *Kalman-Bucy Filters.* Artech House Inc., 1989.

[14] S. A. Brock-Gunn, G. R. Dowling, and T. J. Ellis. Tracking using colour information. Technical Report TCU/CS/1994/7, City University London, 1994.

[15] R. Brooks and S. Iyenga. *Multi-Sensor Fusion: Fundamentals and Applications with Software.* Prentice Hall, 1998.

[16] C. Brown. Tutorial on filtering, restoration, and state estimation. Technical Report TR534, Computer Science Department, University of Rochester, New York, 1995.

[17] C. Brown, H. Durrant-Whyte, J. Leonard, B. Rao, and B. Steer. Distributed data fusion using Kalman filtering: A robotics approach. In M. Abidi and R. Gonzalez, editors, *Data Fusion in Robotics and Machine Intelligence*, chapter 7, pages 267–309. Academic Press Limited, 1992.

[18] L. G. Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376, December 1992.

[19] R. G. Brown. *Introduction to random signal analysis and Kalman filtering.* John Wiley and Sons, 1983.

[20] S. D. Buluswar and B. A. Draper. Color machine vision for autonomous vehicles. *Engineering Applications of Artificial Intelligence*, 11:245–256, 1998.

[21] H. Buxton and S. Gong. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, 78:431–459, 1995.

[22] Q. Cai. *Tracking Human Motion in Indoor Environments Using a Distributed-Camera System.* PhD thesis, University of Texas at Austin, 1997.

[23] Q. Cai and J. K. Aggarwal. Tracking human motion in structured environments using a distributed camera system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(11):1241–1247, 1999.

[24] Q. Cai, A. Mitiche, and J. K. Aggarwal. Tracking human motion in an indoor environment. In *IEEE International Conference on Image Processing*, pages 215–218, Washington, D.C., 1995.

[25] C. Cedras and M. Shah. A survey of motion analysis from moving light displays. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 214–221, 1994.

[26] T. H. Chang and S. Gong. Bayesian modality fusion for tracking multiple people with a multi-camera system. In *2nd European Workshop on Advanced Video-based Surveillance Systems*, London, UK, September 2001.

[27] T. H. Chang and S. Gong. Tracking multiple people with a multi-camera system. In *IEEE Workshop on Multi-Object Tracking*, Vancouver, Canada, July 2001.

[28] T. H. Chang, S. Gong, and E. J. Ong. Tracking multiple people under occlusion using multiple cameras. In *British Machine Vision Conference*, Bristol, England, 2000.

[29] Y. C. Chang and J. F. Reid. RGB calibration for colour image analysis in machine vision. *IEEE Transactions on Image Processing*, 5(10):1414–1422, October 1996.

[30] D. M. Chelberg. *An Approach to Geometric Modeling using Generalized Cylinders and Interpretation of Range Images using Bayesian Networks.* PhD thesis, Stanford University, 1989.

[31] V. Cheng and N. Kehtarnavaz. A smart camera application: DSP-based people detection and tracking. *J. of Electronic imaging*, 9(3):336–346, July 2000.

[32] W. J. Christmas, J. Kittler, and M. Petrou. Structural matching in computer vision using probabilistic relaxation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:749–764, 1995.

[33] S. E. Cohn and D. P. Dee. Observability of discretized partial differential equations. *SIAM J. Numerical Analysis*, 25:586–617, 1988.

[34] R. Collins, A. Lipton, T. Kanadeand, H. Fujiyoshi, D. Duggins, and Y. Tsin. A system for video surveillance and monitoring: VSAM final report. Technical Report CMU-RI-TR-00-12, Carnegie Mellon University, 2000.

[35] I. Cox. A review of statistical data association techniques for motion correspondence. *International Journal of Computer Vision*, 10(1):53–66, 1993.

[36] A. Criminisi, A. Zisserman, L. Van Gool, S. Bramble, and D. Compton. A new approach to obtain height measurements from video. In *Proc. of SPIE*, volume 3576, pages 227–238, November 1998.

[37] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):781–796, August 2000.

[38] R. Cutler, C. Shekhar, B. Burns, R. Chellappa, R. Bolles, and L. Davis. Monitoring human and vehicle activities using airborne video. In *28th Applied Imagery Pattern Recognition Workshop*, Washington, D.C., October 1999.

[39] P. Dagum and R. Chavez. Approximating probabilistic inference in Bayesian networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):246–255, March 1993.

[40] P. Dagum, A. Galper, E. Horvitz, and A. Seiver. Uncertain reasoning and forecasting. *Int. J. of Forecasting*, 11(1):73–87, 1995.

[41] T. Darrell, G. Fordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color and pattern detection. *International Journal of Computer Vision*, 37(2):175–185, 2000.

[42] A. C. Davies, J. Y. Yin, and S. A. Velastin. Computer-based image processing for the monitoring of crowds. *Civil Protection (UK Home Office)*, January 1995.

[43] K. M. Dawson-Howe. Active surveillance using dynamic background subtraction. Technical Report TCD-CS-96-06, Computer Science Department. Trinity College, Dublin, Ireland, 1996.

[44] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*. MIT Press, 1997.

[45] H. Durrant-Whyte. Sensor models and multisensor integration. *International Journal of Robotics Research*, 7(6):97–113, 1988.

[46] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. In *European Conference on Computer Vision*, pages 751–767, 2000.

[47] R. Epstein, A. L. Yuille, and P. N. Belhumeur. Learning object representations from lighting variations. In J. Ponce, A. Zisserman, and M. Hebert, editors, *Object Representation in Computer Vision II*, pages 179–199. Springer-Verlag, 1996.

[48] O. Faugeras. *Three-Dimensional Computer Vision: a Geometric Viewpoint.* MIT Press, 1993.

[49] S. Feyrer and A. Zell. Detection, tracking, and pursuit of humans with an autonomous mobile robot. In *International Conference on Intelligent Robots and Systems*, pages 864–869, Kyongju, Korea, 1999.

[50] N. Friedman and S. Russel. Image segmentation in video sequences: A probabilistic approach. In *Proc. of Conference on Uncertainty in Artificial Intelligence*, 1997.

[51] D. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, January 1999.

[52] T. Gevers and A. Smeulders. Combining color and shape invariant features for image retrieval. *IEEE Transactions on Image Processing*, 9(1):102–119, January 2000.

[53] S. Gong, S. McKenna, and A. Psarrou. *Dynamic Vision: From Images to Face Recognition.* Imperial College Press, World Scientific Publishing, May 2000.

[54] E. Gurewitz, I. Dinstein, and B. Sarusi. More on the benefit of a third eye. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 966–968, 1986.

[55] G. Halevi and D. Weinshall. Motion of disturbances: Detection and tracking of multi-body non-rigid motion. In *IEEE International Conference on Computer Vision and Pattern Recognition*, June 1997.

[56] I. Haritaoglu, D. Harwood, and L. Davis. W4: Who? When? Where? What? A real time system for detecting and tracking people. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 222–227, 1998.

[57] I. Haritaoglu, D. Harwood, and L. Davis. Hydra: Multiple people detection and tracking using silhouettes. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 6–13, 1999.

[58] I. Haritaoglu, D. Harwood, and L. Davis. W4: Real time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830, August 2000.

[59] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[60] A. J. Heap. *Learning Deformable Shape Models for Object*. PhD thesis, University of Leeds, 1997.

[61] D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, November 1996.

[62] G. Helley and Q.-T. Luong. Color in computer vision:. Recent progress. In C. H. Chen, L. F. Pau, and P. Wang, editors, *Handbook of Pattern Recognition and Computer Vision*, pages 283–312. World Scientific, 1998.

[63] S. J. Henkind and M. C. Harrison. An analysis of four uncertainty calculi. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(5):700–714, 1988.

[64] E. C. Hildreth. *The Measurement of Visual Motion*. MIT Press, 1983.

[65] E. Horvitz, J. Breese, D. Heckerman, D. Hovel, and K. Rommelse. The Lumiere Project: Bayesian user modeling for inferring the goals and needs of software users. In *Proc. of the 14th Conf. on Uncertainty in AI*, pages 256–265, 1998.

[66] C. Huang and A. Darwiche. Inference in belief networks: A procedural guide. *International Journal of Approximate Reasoning*, 15(3):225–263, 1996.

[67] T. S. Huang and A. N. Netravali. Motion and structure from feature correspondence: A review. In *Proc. of the IEEE*, volume 82, pages 252–268, February 1994.

[68] Y. J. Huang, H. Dohi, and M. Ishizuka. Man-machine interaction using a vision system with dual viewing angles. *IEICE Transaction on INF. and Syst.*, E80-D(11):1074, 1997.

[69] D. Indyk and S. A. Velastin. Survey of range vision systems. *Mechatronics*, 4(4):417–449, 1994.

[70] S. S. Intille, J. W. Davis, and A. F. Bobick. Real-time closed-world tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 697–703, San Juan, Puerto Rico, June 1997.

[71] M. Isard and A. Blake. CONDENSATION - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 28(1):5–28, 1998.

[72] Y. Ivanov, A. Bobick, and J. Liu. Fast lighting independent background subtraction. *International Journal of Computer Vision*, 37(2):199–207, 2000.

[73] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, January 2000.

[74] R. Jain, R. Kastwi, and B. G. Schunck. *Machine Vision.* McGraw-Hill, 1995.

[75] O. Javed, S. Khan, Z. Rasheed, and M. Shah. Camera handoff: Tracking in multiple uncalibrated stationary cameras. In *IEEE Workshop on Human Motion*, TX, USA, 2000.

[76] F. V. Jensen. *An Introduction to Bayesian Networks.* UCL Press, 1996.

[77] F. V. Jensen and F. Jensen. Optimal junction trees. In *Uncertainty and Artificial Intelligence: Proc. of the Tenth Conference*, 1994.

[78] F. V. Jensen, K. G. Olesen, and S. K. Andersen. An algebra of Bayesian belief universes for knowledge-based systems. *Networks*, 20:637–660, 1990.

[79] G. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *The Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, 1995.

[80] G. A. Jones. Constraint, optimisation and hierarchy: Reviewing stereoscopic correspondence of complex features. *Computer Vision and Image Understanding*, 65(1):57–78, January 1997.

[81] G. A. Jones and P. R. Giaccone. Hierarchical tracking of motion in multiple images. In *IEE Colloquium on Multi-resolution Modelling and Analysis in Image Processing and Computer Vision*, pages 7/1 – 7/6. IEE, London, April 1995.

[82] M. I. Jordan. *Learning in Graphical Models.* Kluwer Academic Press, 1998.

[83] I. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 81–87, 1996.

[84] K. Kanatani. *Geometric Computation for Machine Vision.* Clarendon Press, 1993.

[85] S. B. Kang and R. Szeliski. 3-D scene data recovery using omnidirectional multibaseline stereo. In *IEEE International Conference on Computer Vision and Pattern Recognition*, San Francisco, California, 1996.

[86] P. Kelly, A. Katkere, D. Kuramura, S. Moezzi, S. Chatterjee, and R. Jain. An architecture for multiple perspective interactive video. In *Proc. of ACM Multimedia*, pages 201–212, 1995.

[87] S. Khan and M. Shah. Tracking people in presence of occlusion. In *Asian Conference on Computer Vision*, 2000.

[88] K. Konolige. Small vision system: Hardware and implementation. In *Eighth International Symposium on Robotics Research*, Japan, October 1997.

[89] D. Kortenkamp, M. Huber, F. Koss, W. Belding, J. Lee, A. Wu, C. Bidlack, and S. Rodgers. Mobile robot exploration and navigation of indoor spaces using sonar and vision. In *Proc. of Conference on Intelligent Robotics in Field, Factory, Service, and Space*, pages 509–519, Houston, TX USA, 1994.

[90] J. Krumm and G. Kirk. Video occupant detection for airbag deployment. In *IEEE Workshop on Applications of Computer Vision*, New Jersey, USA, 1998.

[91] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50(2):157–224, 1988.

[92] L. Lee, R. Romano, and G. Stein. Monitoring activities from multiple video streams: Establishing a common coordinate frame. *IEEE Transactions on Pat-*

tern Analysis and Machine Intelligence, Special Issue on Video Surveillance and Monitoring:758–767, 2000.

[93] A. Leon-Garcia. *Probability and Random Processes for Electrical Engineering.* Addison-Wesley, 1989.

[94] J. Z. Li, M. T. Ozsu, and D. Szafron. Modelling of moving objects in a video database. In *International Conference on Multimedia Computing and Systems*, pages 336–343, Ottawa, Canada, June 1997.

[95] Y. Li, S. Gong, and H. Liddell. Recognising the dynamics of faces across multiple views. In *British Machine Vision Conference*, pages 242–251, Bristol, England, 2000.

[96] A. Lipton, H. Fujiyoshi, and R. Patil. Moving target classification and tracking from real-time video. In *Proc. IEEE Image Understanding Workshop*, pages 129–136, 1998.

[97] R. Luo and M. Kay. Data fusion and sensor integration: State-of-the-art 1990s. In M. Abidi and R. Gonzalez, editors, *Data Fusion in Robotics and Machine Intelligence*, chapter 2. Academic Press, 1992.

[98] R. C. Luo and M. G. Kay. *Multisensor integration and fusion for intelligent machines and system.* Ablex Publishing Corporation, 1995.

[99] Q. T. Luong and O. Faugeras. The fundamental matrix: Theory and algorithms, and stability analysis. *International Journal of Computer Vision*, 17(1):43–75, 1996.

[100] C. Mandal, H. Zhao, B. C. Vemuri, and J. K. Aggarwal. 3D shape reconstruction from multiple views. In A. Bovik, editor, *Handbook of Image and Video Processing.* Academic Press, 2000.

[101] M. Marengoni, B. Draper, A. Hanson, and R. Sitaraman. Placing observers to cover a polyhedral terrain in polynomial time. In *Proc. of IEEE Workshop on Applications of Computer vision*, FL USA, 1996.

[102] P. S. Maybeck. *Stochastic Models, Estimation, and Control*, volume 1. Academic Press, 1979.

[103] S. McKenna, S. Gong, and H. Liddell. Real-time tracking for an integrated face recognition system. In *Second European Workshop on Parallel Modelling of Neural Operators*, Faro, Portugal, November 1995.

[104] S. McKenna, S. Gong, and Y. Raja. Face recognition in dynamic scenes. In *British Machine Vision Conference*, 1997.

[105] M. Meyers, T. Ohmacht, and R. Bosch. Video surveillance applications using multiple views of a scene. *IEEE AES Systems Magazine*, pages 13–18, March 1999.

[106] R. Mohr. Projective geometry and computer vision. In C. H. Chen, L. F. Pau, and S. P. Wang, editors, *Handbook of Pattern Recognition and Computer Vision*, pages 313–337. World Scientific Publishing, 1998.

[107] J. M. M. Montiel and L. Montano. Efficient validation of matching hypotheses using Mahalanobis distance. *Engineering Applications of Artificial Intelligence*, 1(3):439–448, June 1998.

[108] R. J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley, 1982.

[109] J. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision*. MIT Press, 1992.

[110] K. P. Murphy. Bayesian map learning in dynamic environments. In S. A. Solla, T. K. Leen, and K. R. Muller, editors, *Advanced in Neural Information Processing Systems 12*, pages 1015–1021. MIT Press, 2000.

[111] S. K. Naya. Catadioptric omnidirectional camera. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 482–488, 1997.

[112] K. C. Ng, H. Ishiguro, M. Trivedi, and T. Sogo. Monitoring dynamically changing environments by ubiquitous vision system. In *IEEE Workshop on Visual Surveillance*, pages 67–73, 1999.

[113] N. Oliver, A. Pentland, and F. Berard. Lafter:A real-time lips and face tracker with facial expression recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1997.

[114] B. D. Olsen. Robot navigation using a sensor network. Master's thesis, Laboratory of Image Analysis, Aalborg University, Denmark, 1998.

[115] T. Olson and F. Brill. Moving object detection and event recognition algorithms for smart cameras. In *Proc. DARPA Image Understanding Workshop*, pages 159–176, May 1997.

[116] J. Orwell, P. Remagnino, and G. A. Jones. Multi-camera colour tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, Colorado USA, 1999.

[117] C. Papageorgiou and T. Poggio. A pattern classification approach to dynamical object detection. In *IEEE International Conference on Computer Vision*, pages 1223–1228, 1999.

[118] C. Papageorgiou and T. Poggio. Trainable pedestrian detection. In *IEEE International Conference on Image Processing*, pages 25–28, 1999.

[119] V. Pavlovic. *Dynamic Bayesian Networks for Information Fusion with Applications to Human-Computer Interfaces.* PhD thesis, University of Illinois, Urbana, IL, USA, 1999.

[120] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA, 1988.

[121] P. Peixoto, J. Batista, and H. Araujo. Real-time human activity monitoring exploring multiple vision sensors. In *Int. Workshop on Intelligent Robotics Systems*, July 1999.

[122] A. Pentland. Classification by clustering. In *Proc. Symposium on Machine Processing of Remotely Sensed Data.* IEEE Computer Society Press, June 1976.

[123] A. Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):107–119, January 2000.

[124] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, 2nd edition, 1992.

[125] P. Pritchette and A. Zisserman. Wide baseline stereo matching. In *ICCV*, pages 863–869, Bombay, India, 1998.

[126] G. Provan, P. Langley, and T. Binford. Probabilistic learning of three-dimensional object models. In *Proc. of the Image Understanding Workshop*, pages 1403–1413, San Francisco, CA. USA, 1996.

[127] K. Rangarajan and M. Shah. Establishing motion correspondence. *CVGIP: Image Understanding*, 54:56–73, 1991.

[128] B. Rao and H. Durrant-Whyte. A fully decentralized multi-sensor system for tracking and surveillance. Technical Report OUEL 1886/91, Oxford University Robotics Research Group, 1991.

[129] B. S. Rao and H. Durrant-Whyte. A decentralized Bayesian algorithm for identification of tracked targets. *IEEE Trans. Systems, Man, and Cybernetics*, 23(6):1683–1698, 1993.

[130] B. S. Y. Rao. Data association methods for tracking systems. In *Active Vision*, pages 91–105. MIT Press, Cambridge, MA, 1992.

[131] D. Raviv and M. Herman. Towards an understanding of camera fixation. In *Proc. International Conference Robotics and Automation*, May 1990.

[132] J. M. Rehg, K. P. Murphy, and P. W. Fieguth. Vision-based speaker detection using Bayesian networks. In *IEEE International Conference on Computer Vision and Pattern Recognition*, June 1999.

[133] C. Ridder, O. Munkelt, and H. Kirchner. Adaptive background estimation and foreground detection using Kalman-filtering. In *Proc. ICRAM*, pages 193–199, 1995.

[134] J. M. Roberts, D. J. Mills, D. Charnley, and C. J. Harris. Improved Kalman filter initialisation using neurofuzzy estimation. In *4th IEE International Conference on Artificial Neural Networks*, 1994.

[135] R. Rosales and S. Sclaroff. Improved tracking of multiple humans with trajectory prediction and occlusion modeling. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1998.

[136] P. L. Rosin and T. Ellis. Image difference threshold strategies and shadow detection. In *British Machine Vision Conference*, pages 347–356, 1995.

[137] R. Rosipal and M. Girolami. An adaptive support vector regression filter: A signal detection application. In *International Conference on Artificial Neural Networks*, volume 2, pages 603–607, Edinburgh, Scotland, 1999.

[138] M. Rossi and A. Bozzoli. Tracking and counting people. In *IEEE International Conference on Image Processing*, volume 3, pages 212–216, Austin, Texas, November 1994.

[139] S. Sakar and K. L. Boyer. Integration, inference, and management of spatial information using Bayesian networks: Perceptual organization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):256–274, 1991.

[140] K. Sato, T. Maeda, H. Kato, and S. Inokuchi. CAD-based object tracking with distributed monocular camera for security monitoring. In *2nd CAD-Based Vision Workshop*, pages 291–297, Champion PA. USA, 1994.

[141] C. Schmid and A. Zisserman. Automatic line matching across views. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 666–671, 1997.

[142] M. Sekiguchi, H. Okada, and N. Watanabe. Neural network based landmark detection for mobile robot. In *Proc. of SPIE*, volume 2760, pages 216–223, 1996.

[143] N. Sgouros, G. Papakonstantinou, and P. Tsanakas. Localized qualitative navigation for indoor environments. In *IEEE Conference on Robotics and Automation*, Minneapolis, USA, 1996.

[144] R. Shachter. Evaluating influence diagrams. *Operations Research*, 34:871–882, 1986.

[145] S. Shahand, J. Eledath, J. Ghosh, and J. K. Aggarwal. Multisensor integration for scene analysis: An experiment in human form detection. In *IEEE International Conference on Image Processing*, volume 2, pages 199–202, Santa Barbara, CA, 1997.

[146] L. S. Shapiro. *Affine Analysis of Image Sequences.* PhD thesis, University of Oxford, 1993.

[147] L. S. Shapiro, A. Zisserman, and M. Brady. 3D motion recovery via affine epipolar geometry. *International Journal of Computer Vision*, 16(2):147–182, 1995.

[148] J. Sherrah and S. Gong. Fusion of perceptual cues using covariance estimation. In *British Machine Vision Conference*, Nottingham, England, September 1999.

[149] J. Sherrah and S. Gong. Resolving visual uncertainty and occlusion through probabilistic reasoning. In *British Machine Vision Conference*, Bristol, England, September 2000.

[150] Y. Shirai. *Three-Dimensional Computer Vision*. Springer-Verlag, 1987.

[151] A. Singhal and C. Brown. Dynamic Bayes net approach to multimodal sensor fusion. In *Proc. of SPIE*, volume 3209, October 1997.

[152] A. R. Smith. Color gamut transform pairs. In *SIGGRAPH*, pages 12–19, 1978.

[153] A. Smola, B. Scholkopf, and K.-R. Muller. General cost functions for support vector regression. In T. Downs, M. Frean, and M. Gallagher, editors, *Proc. of the Ninth Australian Conf. on Neural Networks*, pages 79–83, Brisbane, Australia, 1998.

[154] A. J. Smola and B. Scholkopf. A tutorial on support vector regression. Technical Report NC2-TR-1998-030, Royal Holloway College, London, UK, October 1998.

[155] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis and Machine Vision*. International Thomson Publishing Company, 1993.

[156] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, 1999.

[157] M. Stevens and J. Beveridge. Using multisensor occlusion reasoning in object recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1997.

[158] S. Stillman, R. Tanawongsuwan, and I. Essa. A system for tracking and recognising multiple people with multiple cameras. Technical report, Georgia Institute of Technology, 1998. Technical Report No. GIT-GVU-98-25.

[159] T. M. Strat and M. A. Fischler. Context based vision: Recognizing objects using information from both 2-D and 3-D imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1050–1065, 1991.

[160] M. Swain and D. Ballard. Index via color histogram. In *IEEE International Conference on Computer Vision*, pages 390–393, 1990.

[161] R. Talluri and J. K. Aggarwal. Position estimation techniques for an autonomous mobile robot - A review. In C. H. Chen, L. F. Pau, and P. Wang, editors, *Handbook of Pattern Recognition and Computer Vision*, chapter 4.4, pages 765–796. World Scientific, 2nd edition, 1998.

[162] K. Tarabanis, P. Allen, and R. Tsai. A survey of sensor planning in computer vision. *IEEE Transactions on Robotics and Automation*, 11(1):86–104, February 1995.

[163] J. Taylor, T. Olson, and W. N. Martin. Accurate vergence control in complex scenes. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 540–545, 1994.

[164] M. Teal and T. J. Ellis. Spatial temporal reasoning based on object motion. In *British Machine Vision Conference*, pages 465–474, Edinburgh, UK, 1996.

[165] E. Thirion and C. Ronse. Self calibration and 3D reconstruction from lines with a single translating camera. In *British Machine Vision Conference*, 1996.

[166] C. Tomasi and R. Manduchi. Stereo matching as a nearest-neighbor problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):333–340, 1998.

[167] P. Torr, A. Fitzgibbon, and A. Zisserman. Maintaining multiple motion model hypotheses over many views to recover matching and structure. In *IEEE International Conference on Computer Vision*, pages 485–491, January 1998.

[168] P. Torr, T. Wong, D. Murray, and A. Zisserman. Cooperating motion processes. In *British Machine Vision Conference*, 1991.

[169] P. Torr and A. Zisserman. Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15:591–605, 1997.

[170] G. T. Toussaint. Computational geometry and computer vision. In R. A. Melter, A. Rosenfeld, and P. Bhattacharya, editors, *Vision Geometry, Contemporary Mathematics*, volume 119, pages 213–224. American Mathematical Society, 1991.

[171] K. Toyama and E. Horvitz. Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *Asian Conference on Computer Vision*, Taipei, Taiwan, January 2000.

[172] R. Y. Tsai. A versatile camera calibration technique for high accuracy 3D machine vision metrology using off the shelf TV cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323–344, August 1987.

[173] R. Y. Tsai and T. S. Huang. Estimating three-dimensional motion parameters of a rigid planar patch. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 29(6), December 1981.

[174] S. Ullman. *High-level Vision.* MIT Press, Cambridge, MA, 1996.

[175] R. Vanderbei. Loqo: An interior point code for quadratic programming. Technical report, Princeton University, 1994. Technical Report SOR 94-15.

[176] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, New York, 1995.

[177] C. J. Veenman, M. Reinders, and E. Backer. A composite model and algorithm for motion correspondence. In *Proc. of the Sixth Annual Conference of the Advanced School for Computing and Imaging*, Belgium, June 2000.

[178] J. Verestoy and D. Chetverikov. Experimental comparative evaluation of feature point tracking algorithms. In *Proc. 22nd Workshop of the Austrian Pattern Recognition Group*, pages 255–263, Illmitz, Austria, 1998.

[179] R. L. Vergnet, S. B. Pollard, and J. E. W. Mayhew. Stereo-matching of line segments based on a 3-dimensional heuristic with potential for parallel implementation. In *Proc. of the Alvey Vision Conference*, pages 181–186, England, September 1989.

[180] E. Walker, M. Herman, and T. Kanade. A framework for representing and reasoning about three-dimensional objects for vision. In S. S. Chen, editor, *Advances in Spatial Reasoning*, volume 1, chapter 6, pages 219–247. Ablex Publishing, 1990.

[181] C. Wren, A. Azerbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:780–785, 1997.

[182] G. Xu and Z. Zhang. *Epipolar Geometry in Stereo, Motion and Object Recognition: A Unified Approach.* Kluwer Academic Publishers, 1996.

[183] T. Yamane, Y. Shirai, and J. Miura. Person tracking by integrating optical flow and uniform brightness regions. In *IEEE International Conference on Robotics and Automation*, pages 3267–3272, 1998.

[184] Z. F. Yang and W. H. Tsai. Using parallel line information for vision-based landmark location estimation and an application to automatic helicopter landing. *Robotics and Computer-Integrated Manufacturing*, 14:297–306, 1998.

[185] R. Yogesh, S. J. Mckenna, and S. Gong. Segmentation and tracking using colour mixture models. In *Asian Conference on Computer Vision*, pages 607–614, Hong Kong, 1998.

[186] K. C. Yow and R. Cipolla. Feature-based human face detection. *Image and Vision Computing*, 15(9):713–735, 1997.

[187] K. Yoyama, J. Krumm, B. Brumitt, and B. Meyes. Wallflower: Principles and practice of background maintenance. In *IEEE International Conference on Computer Vision*, 1999.

[188] T. Yuan and M. Subbarao. Integration of multiple-baseline color stereo vision with focus and defocus analysis for 3D shape measurement. In *Proc. of SPIE*, volume 3520, 1998.

[189] Z. Zhang. Understanding the relationship between the optimization criteria in two-view motion analysis. In *IEEE International Conference on Computer Vision*, pages 772–777, 1998.

[190] Z. Zhang and O. Faugeras. Three-dimensional motion computation and object segmentation in a long sequence of stereo frames. *International Journal of Computer Vision*, 7(3):211–241, 1992.

[191] Z. Y. Zhang. Parameter estimation techniques: A tutorial with respect to conic fitting. *Image and Vision Computing*, 15(1):59–76, 1996.

[192] L. Zhao and C. Thorpe. Stereo- and neural network-based pedestrian detection. In *Proc. of the IEEE Intelligent Transportation Systems Conference*, 1999.