PEDESTRIAN TRACKING BY LEARNING OF MULTI-VIEW HUMAN COLOR APPEARANCE

By Huynh Loc Huu (1010214)

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Associate Professor Kazunori KOTANI

September, 2012

PEDESTRIAN TRACKING BY LEARNING OF MULTI-VIEW HUMAN COLOR APPEARANCE

By Huynh Loc Huu (1010214)

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Associate Professor Kazunori KOTANI

and approved by
Associate Professor Kazunori KOTANI
Associate Professor Atsuo YOSHITAKA
Professor Jainwu DANG

August, 2012 (Submitted)

Abstract

Recent years, cameras have been also widely applied in varying folds of life according to low production costs. However, these surveillance systems have not already reached the expected performance yet because of lacking the specialized software. Therefore, in this thesis, we aim to build the pedestrian tracking system which can reliably detect and tracking the unknown number of walkers by using multiple cameras. The system has to keep in track multiple targets even in the urban environments where clutter and occlusion occur frequently. Results from our automated tracking system may provide helpful insights for evacuation planning and for real-time situation awareness during the emergency response to public disturbances.

We start by performing background subtraction independently on every available view. Besides using the blobs tracking from the result of background subtraction, we also use the color appearance. We try to learn the color appearance of pedestrian while they move inside the tracking area. When people are moving closely to another one, or they are occluded for a long time, the identity switching cases will occur easily. However, with color appearance learning, it not only helps to overcome such that problem, but the accuracy of localization is also better than methods, which use the background subtraction only.

In some previous methods, a person is modeled as the cylinder which is projected by rectangles in each view. However, the rectangle did not enclose the foreground pixels of people very accurately, and the color appearance cues may be affected by ambiguous pixels from the background. In order to improve the result of tracking, we defined the new model which is the cooperation between the rectangle and the ellipse.

We also proposed a detection and tracking in one-step method based on a modified Bayesian model. At each frame, the marginal conditional probability of the state of all people is approximated based on the background subtraction images from all views and the color appearance model. The marginal conditional probability in one frame only depends on one previous frame. This implementation helps to reduce the searching space which is the problems of the original Bayesian model which try to connect all previous frames. The approximation is obtained by sampling the posterior function using Reversible Jump Markov Chain Monte Carlo method. This system is really general, which means that besides the background subtraction and color appearance, other higher features could be applied to improve the accuracy, for example Histogram of Orient (HOG).

Acknowledgment

This dissertation would not have been possible without the assistance and help of a great number of people. For their love, support, and professional guidance, I would like to gratefully acknowledge the following people.

First and foremost I offer my sincerest gratitude to my supervisor, Professor Kotani Kazunori, who has supported me throughout my study with his knowledge, guidance and encouragement. His dedication for research and curiousness had always been an inspiration; his patience, belief in my abilities, and incredible depth of knowledge provided invaluable support for me through the tough times. Without his consistent help, this thesis would not have been completed or written. I would also like to thank Chen Fan for his support both with my dissertation and beyond.

In addition, I would like to thank the Japanese Government (Monbukagakusho) Scholarship Program for financial support during my stay in Japan.

Last but not least, I thank my family who endured this long process with me, always offering love, support and understanding. Thanks are also due to numerous friends, especially those at Kotani Laboratory for their willingness to participate in challenging discussion and give help to tackle the language barrier in my daily life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent years, the strong development of computational speed makes it possible to implement many complex algorithms in real time. In addition, cameras have been also widely applied in varying folds of life according to low production costs. The real state points out how usefully cameras are used in different places, ranging from the cell phones, laptops to cars, elevators, airplanes or houses. Therefore, the field of computer vision has necessary conditions to become one of the most important researches recently.

In the common trend, the application of surveillance cameras also proves its serviceable advances in urban environments, where the number of dangerous events steadily increases. These cameras are primarily located in potentially crowded areas, such as airports, stations, shopping malls, stadiums or tourist attractions. They provided realistic information by color videos for road monitoring or preventing the crime inside the cities.

However, these surveillance systems have not already reached the expected performance yet because of lacking the specialized software. Almost of monitoring tasks are passively observed by human, only a small fraction is automatically active, such as traffic speeding reporting application. Therefore, improving the effectiveness of using cameras and reducing the workload of human by intelligent software become one of the most notable researches in the field of computer vision. In this context, the ultimate goal is to build a smart system which can observe the urban centers crowded with people and car, detect and identify tragic events, potential accidents or abnormal threats, and report them automatically to the security authority. It is also the primary objective of this thesis. Particularly, we aim to build the pedestrian tracking system which can reliably detect and tracking the unknown number of walkers by using multiple cameras.

1.1 Thesis goal

Despite useful conditions such as the fast speed of computation and the high resolution of cameras, implementing such an smart surveillance system is very difficult and raising a lot of challenges.

- First of all, detecting and tracking pedestrians in video of a crowded scene is a challenging problem due to differences in deformable appearances. The urban environments are frequently cluttered with obstacles that may occlude, reflect, or having similar appearance on the target's objects. The spatial overlap between people makes it difficult to distinguish individuals.
- Particularly, besides of moving by group, pedestrians also often perform many kinds of complicated motions and interact with others. Therefore, the tracking task becomes more complex by these activities.
- In addition, the variety of behaviors makes it really hard to distinguish between normal and abnormal events. So that, the task of detecting and identifying tragic events, potential accidents or unusual threats, is one of the most complex fields in research.

Therefore, the thesis is motivated by these challenges and the pitfalls which existing methods encountered. We aim to build the intelligent surveillance system by using a Bayesian model and mixing the detection and tracking in one task. The ultimate goal of this system is the ability to detect and track a varying number of people by using multiple cameras (see Figure 1.1). The result in the figure presents the locations of all pedestrians in the tracking areas. Each people are bounded by different rectangle. More specifically, the system needs to address three key problems:

- Reliably detecting and tracking pedestrians under reasonable crowd densities and from different viewpoints. The system has to keep in track multiple targets even in the urban environments where clutter and occlusion occur frequently.
- The model should be as comprehensive as possible in order to flexibly integrate other specific models and useful clues. As more general the model is, the system could be used in many applications instead of only tracking pedestrians, such as detecting the event of people falling down on the floor.
- The tendency of moving such as spatial location and velocity can be used to smooth the tracking trajectories.

Combining these requirements in a robust framework will create the intelligent

system which is capable of automatically monitoring the pedestrians in public areas as well as contributing a new methodology to the research of social behavior.

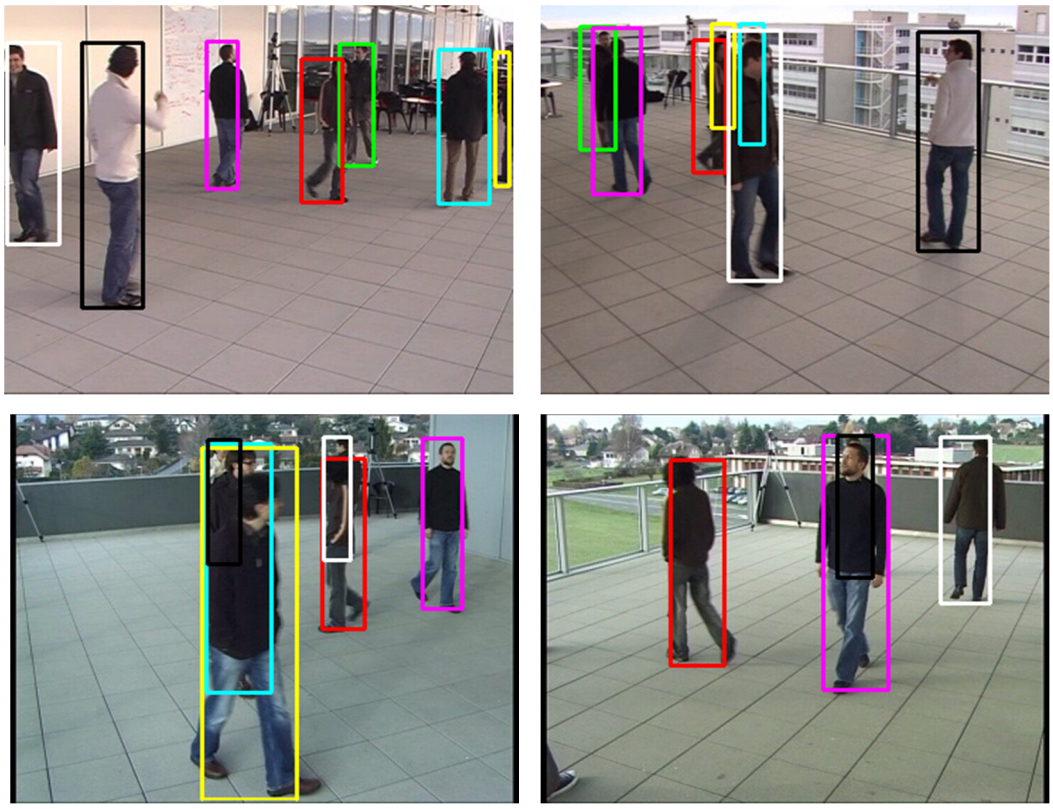

Figure 1.1 Tracking result by using multiple cameras

The final system in this thesis showed its usefulness by being embedded in several potential applications.

- Pedestrian surveillance is the obvious one. The available function is only the ability to track trajectories of objects; therefore, it has not information enough to completely understand complex situations involving humans. However, with the sufficient knowledge about motion of pedestrian, it also makes the system useful in the several public areas where the main body motions are simple, such as a passageway, hall or corridor. It will help to reduce the human workload of security authority by detecting and identify tragic events, potential accidents or abnormal threats, and report them automatically to the security center.
- Furthermore, in the research of analysis customer behavior, the trajectories' information is especially helpful to increase the amount of sales. It can navigate the interesting location of the customer in market, and their paths also point out how they moved to shopping. By this

system, businessman can rearrange their retail location, which will increase the attractiveness and convenience to the customer.

- Another example is analyzing team sports' tactical information, for example, in football or basketball. The knowledge about the players' actions and moving trajectories is very important for coaches to evaluate the performance. It could also be used to analyze the opposing teams' tactics. Instead of wasting hours to watching video and finding out which tactic the opposite teams played, coaches can benefit from the automatic system which analyzed all statistical data very fast and correctly

In order to capably serve above complex tasks, a tracking system must be robust enough to keep in tracking people accurately and smoothly with a minimum amount of identity switches, miss-detections and false positives, even not in suboptimal conditions. Precision is a solid characteristic for gathering significant statistical data about people motion. In addition, the system needs an ability to extend flexibly the goal of application. For example, it should be easy to enlarge the searching space which cannot only track pedestrian but also detecting the emergency of falling down to the floor. The implementation of such a system is the objective of this thesis, which we define more detail in the rest of this chapter.

1.2 Related work

In this section, we present previous works and fundamental backgrounds in the field of pedestrian tracking research. Generally, the literature on tracking multiple targets in cluttered scenes can be divided into two categories: monocular approaches and multi-view approaches. The state of the art can be referred in the recent survey by Yilmaz et al [1].

1.2.1 Monocular Approaches

First of all, in history of multiple targets detecting and tracking research, there are many algorithms based on the single camera. Although this approach has its advantage of simple installation, lacking of 3D information makes it difficult to deal with denser situations where occlusions occur more frequently.

Blob tracking is a popular low cost approach for tracking objects [2], [3]. It

entails extracting blobs in each frame, and tracking is performed by associating blobs from one frame to the next. The BraMBLev system [4], for example, is a multi-blob tracker that generates a blob-likelihood based on a known background model and appearance models of the tracked people. Its performance degrades when multiple objects merge into one blob due to proximity or occlusions. Alternate approaches maintain explicit object states with position, appearance, and shape. Zhao and Nevatia [4], [6] present interesting results when tracking multiple people with a single camera. They use articulated ellipsoids to model human shape, color histograms to model different people's appearance, and an augmented Gaussian distribution to model the background for segmentation. Once moving head pixels are detected in the scene, a principled MCMC approach is used to maximize the posterior probability of a multi person configuration. This concept of global trajectory optimization was previously explored in [7] and more recently in [8]. It also forms the basis of our tracking formulation; however, there is an important difference. Our approach utilizes fusion of multiple views and combines the task of detection and tracking seamlessly.

Okuma et al. [9] propose a noteworthy combination of Adaboost for object detection and particle filters for multiple objects tracking. The combination of these two approaches leads to fewer failures than either one on its own, as well as addressing both detection and consistent track formation in the same framework. Brostow and Cipolla [10] present a probabilistic framework for the clustering of feature point trajectories to detect individual pedestrians in a crowd. These and other similar approaches like [11], [12], [13] skip the modeling of articulations in favor of appearance models trained for specific un-occluded views of their respective subjects. As a result, they are challenged by fully and partially occluding objects, as well as appearance changes.

A number of monocular tracking techniques have been devised for handling occlusions. The typical approach is to detect the occurrence of occlusion by blob merger [2]. The methods for tracking feature points simply detect the occlusion of a feature point as the disappearance of the point being tracked [15]. In recent years, tracking techniques using object contours [16], [17] and appearances [18], [19], which represent and estimate occlusion relationships between objects by using hidden variables of depth ordering of objects toward the camera, have been proposed. Wu et al. [18] incorporate an additional hidden process for occlusion into a dynamic Bayesian network and rely on the statistical inference of the hidden process to reveal occlusion relations. Senior et al. [20] use appearance models to

localize objects and use disputed pixels to resolve their depth ordering during occlusions. However, the system cannot maintain object identity after occlusions. Jojic and Frey [21] and Tao et al. [22] both model videos as a layered composition of objects and use EM to infer object's appearances and motions. Recently, Perera et al. [23] proposed a two stages framework, which involves one to one correspondence followed by a split and merge analysis, or linking tracks across occlusions.

Most of the aforementioned approaches rely on partial observations, which make it difficult to handle full occlusions. In addition, small and consistent motions are assumed to predict the motion patterns of objects through occluded views. This causes problems in dealing with long periods of occlusions of an object under unpredictable motions. In spite of the current body of knowledge, we believe monocular methods have limited ability to handle occlusions involving several objects, generally two or three, because the single viewpoint is intrinsically unable to observe the hidden areas.

1.2.2 Multi-camera Approaches

The use of multiple cameras soon becomes necessary when one wishes to accurately detect and track multiple occluding people and compute their precise locations in a complex environment. Multi-view tracking techniques intend to decrease the hidden regions and provide 3D information about the objects and the scene by making use of redundant information from different viewpoints.

In [24], Kelly et al. constructed a 3D environment model using the voxel feature. Humans were modeled as a collection of these voxels to resolve the camera-handoff problem. In [25], Sato et al. use CAD-based environment models to extract 3D locations of unknown moving objects. Jain and Wakimoto [26] also utilized calibrated cameras to obtain 3D locations of each object in an environment model for the Multiple Perspective Interactive Video. These works were characterized using environment models and calibrated cameras. Multi targets tracking by association across multiple views was addressed in a series of papers from the latter half of the 1990s. In [27], Nakazawa et al. constructed a state transition map that linked regions observed by one or more cameras, along with a number of action rules to consolidate information between cameras. Orwell et al. [28] present a tracking algorithm to track multiple objects in multiple views using "color" tracking. They model the connected blobs obtained from background subtraction using color histogram techniques and use them to match and track objects. Cai and Aggarwal [29] extend a single-camera tracking system by starting with tracking in a single

camera view and switching to another camera when the system predicts that the current camera will no longer have a good view of the subject. Spatial matching was based on the Euclidean distance of a point to its corresponding epipolar line. In [30], individuals are tracked both in image planes and top view using a combination of appearance and motion models. Bayesian networks were used in several papers as well. In [31], Chang and Gong used Bayesian networks to combine geometry (epipolar geometry, homographic, and landmarks) and recognition (height and appearance) based modalities to match objects across multiple sequences. Bayesian networks were also used by Dockstader and Tekalp in [32], to track objects and resolve occlusions across multiple calibrated cameras. Integration of stereo pairs is another popular approach, adopted by [33], [34], [35], [36] among others. Krumm et al. [33] use stereo cameras and combine information from multiple stereo cameras in 3D space. They perform background subtraction and then detect human-shaped blobs in 3D space. Color histograms are created for each person and are used to identify and track people. Mittal and Larry [34] use a similar method to combine information in pairs of stereo cameras. Regions in different views are compared with each other, and back projection in 3D space is done in a manner that yields 3D points guaranteed to lie inside the objects.

Although these methods attempt to resolve occlusions, the underlying problem of using features (appearance templates, blob shapes) that might be corrupted due to occlusions remains. Second, occlusion reasoning in these approaches is typically based on temporal consistency in terms of a motion model, whether it is Kalman filtering or more general Markov models. As a result, these approaches may not be able to recover if the process begins to diverge. As well as cases of near total occlusion, the people are dressed in very similar colors. Using blob shapes or color distributions for region matching across cameras may lead to incorrect segmentations and detections.

The homographic occupancy constraint [37] presented in this paper fuses information from multiple views using sound geometrical constructs and resolves occlusions by localizing people on multiple scene planes. We essentially attempt to find image locations of scene points that are guaranteed to be occupied by people. These occupancies are then used to resolve occlusions and track multiple people. In this context, the work by Mittal and Larry [34], Franco and Boyer [38], Berclaz et al. [8], Yang et al. [39], and the parallel work on range sensor based occupancy grids for robot navigation is quite relevant [40], [41]. However, our approach is also motivated by these approaches, which indirectly fuse information in 3D space from

calibrated cameras. Moreover, all of evidences are gathered from all the cameras into a unified synergistic framework where occlusion resolution, detection, and tracking are performed simultaneously. The detection and tracking results are then propagated back to each view.

1.2.3 Approaches of Multi-targets tracking

Despite of monocular or multi-view approaches, multi targets tracking is also one of the most important parts. Object tracking has a long history in computer vision. Compared to tracking a single target, multi-targets tracking is a lot more complicated: the single target case can be solved by detecting the object in each frame possibly only within a local region around the predicted position and "connecting the dots" to a consistent trajectory; for multiple targets the problem is much more complex due to the data association problem, and to interactions between different targets (e.g. inter-object occlusion). An additional difficulty is that in most scenarios the number of targets is not known a priori, and may in fact vary over time. Early work mostly focused on recursive methods, where the current state depends only on the previous one: initially Kalman filtering, e.g. [42], and later particle filtering [43, 9, 45], which represents the posterior by a set of samples rather than an analytic expression, and can thus better cope with ambiguous, multi modal distributions.

Recently, several non-recursive approaches have appeared, which aim to formulate the problem such that a solution can be found which is (in some cases globally) optimal over a longer time interval. One way to reduce the immense solution space of tracking over extended time windows is to commit in advance to a restricted set of possible target locations [46, 47, 48, 49], which are usually found by appearance based object detection [50, 51] or by background subtraction [52]. The tracker is forced to form trajectories through these locations, without taking into account localization uncertainty. A different approach pursued in [53, 8, 55] discretizes the space of possible locations to a regular grid, which avoids early commitment to detection results, but instead introduces discretization errors. The resulting optimization problems are either quadratic integer programs [46, 48], in which case they are solved to local optimality by custom heuristics based on recursive search or graph cuts; or integer linear programs (ILP) [53, 55, 47], which are solved to near-global optimality through LP-relaxation. An exception is [49], which solves a simplified version of the problem without occlusions to global optimality with a network flow algorithm, then greedily adds occluded targets.

Recent research trend thus tries to address the problem by decoupling detection

from tracking. A detector is applied at each time step independently and a data association as above methods links the detections together, producing more robust results. However, because the detection step is executed freely to the tracking step, the state space grows due to large ground, such a problem of [8]. Furthermore, the recursive tracking makes the search space grow exponentially with the number of frames. On the other hand, the optimization methods require a processing on a batch of frames, which cost amount of computation; therefore, it could not be used in the applications which need the result of tracking from frame to frame, such as security surveillance.

In order the reach the good performance as well as applicable in real situation, we proposed a new method which used a modified Bayesian model, which helps to detect and track pedestrian in one step from frame to frame while reducing the searching space. We increase the accuracy of the detection by mixing the foreground mask and the color appearance. The main idea is the color appearance of each object is automatically learned and updated by each frame. We show that this improvement helps to robustly overcome the problem of identities switching. In addition, dynamic model with the velocity reduces the tracking state space while smoothing the trajectories. We will look over the overview of proposed approach below.

1.3 Overview of the approach

First of all, we divide the ground plane into a grid of cells in order to discretize the searching space as in [8]. When the scaling of division is smaller, the result of tracking is more accurate. In [8], the searching space increases due to the larger size of grid because they try to calculate the occupancy probabilities for all positions on the grid. However, it's not the problem of our proposed method because computational complexity is only depended on the number of people. We only find the probabilities for those neighbors around previous position of each people.

We start by performing background subtraction independently on every available view. Besides using the blobs tracking from the result of background subtraction, we also use the color appearance. We try to learn the color appearance of pedestrian while they move inside the tracking area. The learned color histogram is then compared to the observed color histogram from cameras to determine whether that's the same people or not. Because the color scaling of each camera is

different, we compare the color appearance in each view separately, then fusing all the matching result in one final color likelihood function. When people are moving closely to another one, or they are occluded for a long time, the identity switching cases will occur easily. However, with color appearance learning, it not only helps to overcome such that problem, but the accuracy of localization is also better than methods, which use the background subtraction only.

In some previous methods, a person is modeled as the cylinder which is projected by rectangles in each view. However, the rectangle did not enclose the foreground pixels of people very accurately, and the color appearance cues may be affected by ambiguous pixels from the background. In order to improve the result of tracking, we used the cooperation between the rectangle and the ellipse. The mixing model is used when connecting the statistical information of background subtraction from all views, nonetheless, when comparing and learning the color appearance on each view, the seven parts of 2D ellipse is applied. The seven parts of ellipse could not only enclose the pedestrian better, but it could also mark the interest area on the color appearance.

In this thesis, we proposed a detection and tracking in one-step method based on a modified Bayesian model. At each frame, the marginal conditional probability of the state of all people is approximated based on the background subtraction images from all views and the color appearance model. The marginal conditional probability in one frame only depends on one previous frame. This implementation helps to reduce the searching space which is the problem of the original Bayesian model which tries to connect all previous frames. The approximation is obtained by sampling the posterior function using Reversible Jump Markov Chain Monte Carlo method. This system is really general, which means that besides the background subtraction and color appearance, other higher features could be applied to improve the accuracy, for example Histogram of Orient (HOG). On the other hand, because the state of searching space is not limited only in location of people, the system can be extended to serve other useful functions. For example, by rotating the 2D ellipse, we can determine the rotation angle of people, which can detect the event of falling people. Another advantage of this proposed method is it could not only fuse the information from all multiple cameras; it could also use the color of each view to get a better result. An overview of our approach is illustrated in Figure 2.1.

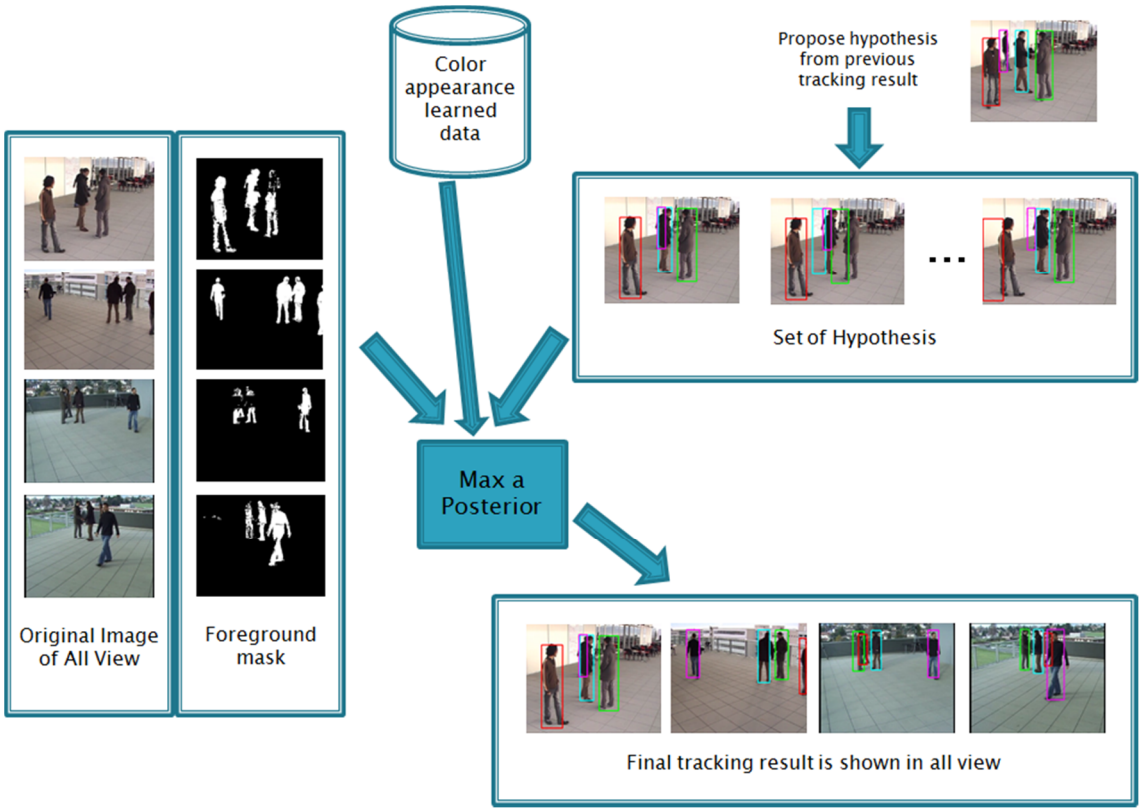

Figure 1.2 Overview of our approach

1.4 Outline

The remainder of this thesis is structured as follows:

- In Chapter 2, we introduce the framework on which our system is built. We also briefly expose some of the methods used by our system, but that were not precisely the focus of this work, such as camera calibration or background subtraction.
- In Chapter 3, we explain our main method for people detection and tracking based on modified Bayesian formulation.
- In Chapter 4, we have an evaluation on two dataset, which could show the accuracy of our system and comparing to the state of art.
- Finally, after some perspectives for future work, we conclude in Chapter 5.

Chapter 2

Framework

In this chapter, we introduce the framework where we build proposed method to detect and track pedestrian. We present the video data which were used for evaluating the performance of our algorithm. We also explain some existed methods used in our system, such as background subtraction, color histogram and EMD matching method. These algorithms are essential elements, though they are not the center of our research. Finally, we present the metrics we used to evaluate the performance of our algorithms.

2.1 Tracking area: Grid Discretization, Multiple Cameras and Cameras Calibration

The first basic framework is the tracking area where cameras are installed to observe pedestrians. Generally, tracking area can be the corridor in some public places such as station, airport, or the pathway in some buildings. Instead of relying on continuous geometric coordinates to locate the detected objects in a common reference plane, we divide the ground plane into a finite number of cells (se Figure 2.1). The advantage of this discretization is that we could implement our method as a discrete problem and increase the speed of the system by preventing the complexity of continuous computation.

Multiple people moving tend to occlude each other, which generates ambiguity when observed by a single camera. When the number of people is small, researchers usually address this issue by relying on time consistency. However, when the people density increases, the large amount of occlusions produced renders any monocular tracking task very difficult. These ambiguities may be eliminated by images taken from another view point. Although this issue can also be partly addressed by using a single top mounted camera that may reduce the amount of occlusions, this solution is not without flaws. Since the useful field of view of any type of camera is limited, using multiple cameras is also a way to expand the surveillance area. Furthermore, it provides more accurate localizations than a single camera setup.

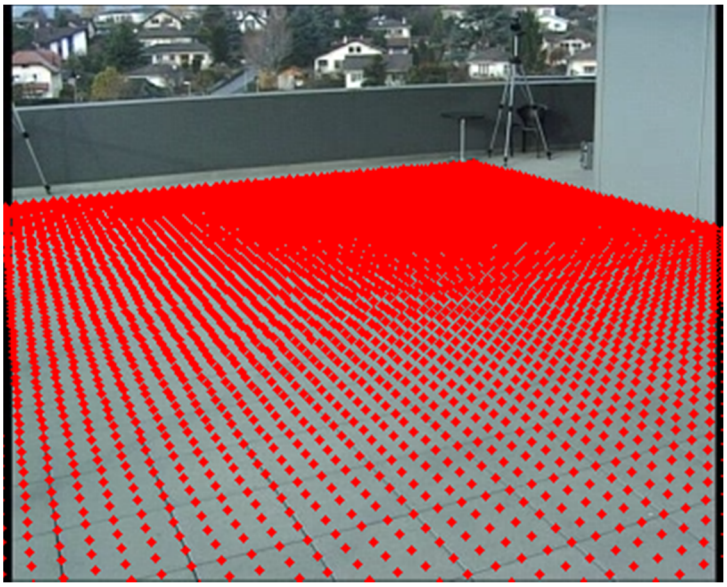

Figure 2.1 Ground plane is discretized into a finite number of locations. The red points are the locations.

In order to make the performance of localization better, we use a high resolution grid in which the distance between two locations is only 12.5 cm. Although the resolution could be increased, but we think it's enough to reach the best performance while reduce the searching space as much as possible.

Multi-view based tracking shows its main strength from the additional information provided by different view-points, which can overcome occlusion. However, this extra knowledge can only be fully exploited if the connection between image measurements and scene measurements is known. In every multi-view approach, detections in individual views eventually need to be related and merged. This correspondence problem is usually dealt with by computing the cameras calibration, which consists, for every camera, in estimating its intrinsic parameters such as focal length, principal point and radial distortion; among others and extrinsic parameters, that is, its position and orientation in space. Although we used some video data, which had been calibrated by the provider, but we guess that the knowledge about calibration is useful for those who want to build their own data. Over the years, several different camera calibration methods [14; 44; 54] have been developed. Their computation typically requires the definition of correspondences between scene and image measurements. A technique called auto calibration [56; 57; 58] allows to trade extensive scene knowledge for knowledge of camera motion. Once estimated, the camera calibration parameters give a precise understanding of the image formation mechanism. A complete discussion of the camera calibration problem extends well beyond the scope of this work, and we refer the interested

reader to [59], which covers extensively the multiple aspects of the subject. The result of calibration is shown in Figure 2.2 where some corresponding points were presented in all views.

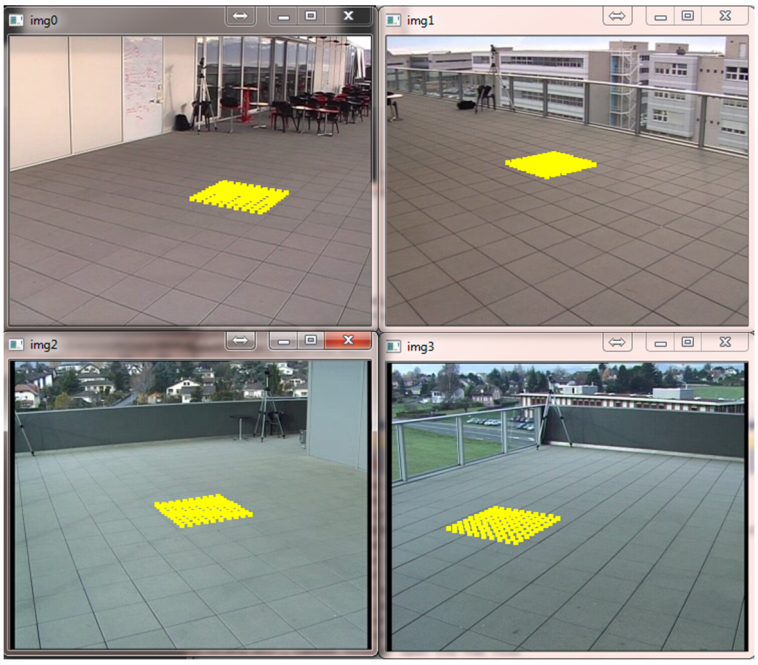

Figure 2.2 Corresponding points in all views show us the result of calibration

2.2 People modeling

We show here a generative model which approximates the appearance of people for background subtraction images and color appearance matching. In order to fit the requirement for fusing information from both subtraction images and color histogram matching, we proposed the model of mixing between rectangle and ellipse.

A human body is a challenging object to model, because it is both highly articulated and deformable. As suggested by the example set of silhouettes in Figure 2.3, pedestrian silhouettes can take very varied shapes. They can be even more heterogeneous when people perform activities different from walking, such as running or playing sports. Therefore, no particular fixed shape can faithfully capture the wide range of potential silhouettes generated by pedestrians, and one has to rely on complex articulated models. However, we realize that usual pedestrian silhouettes share a common shape. More generally, pedestrians occupy a portion of space that can be roughly approximated by two parts. One part is rectangle, and another is ellipse as in Figure 2.4.

We refer to Figure 2.5 for easily realizing the advantages of the model. We have the foreground mask which was subtracted from the background. The white pixel belongs to people who existed in the observing scene. However, in the figure, there is one person who was enclosed by cyan pixels. These cyan pixels are the approximation model for this people. The ellipse could enclose almost the body while the below rectangle enclose two legs regardless to the shape of behavior.

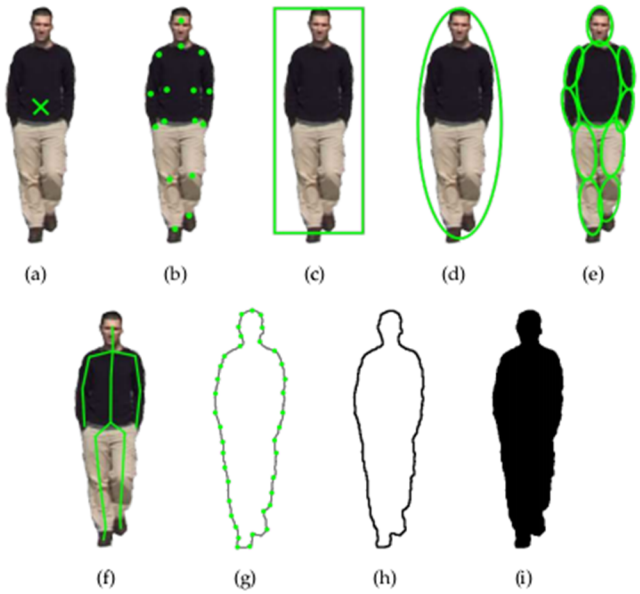

Figure 2.3 Different types of people representation found in the literature: (a) centroid, (b) multiple points, (c) bounding box, (d) ellipse, (e) multiple ellipses, (f) skeleton, (g) control points, (h) contour and (i) silhouette

Figure 2.4 The rectangle and ellipse mixing human model

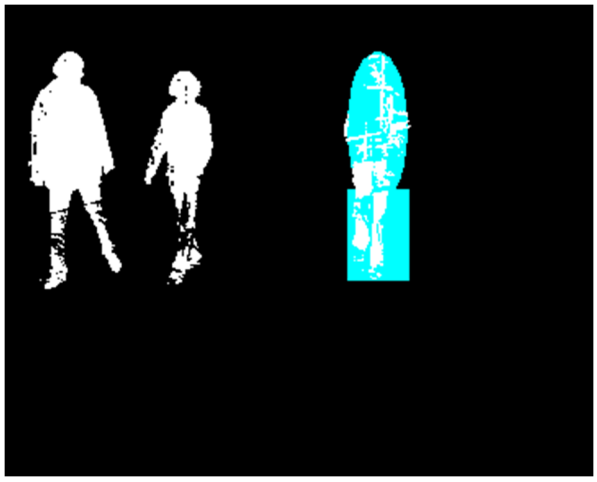

Figure 2.5 The advantage of our human model for approximating the foreground mask

The cyan pixels which belong to the ellipse become the mask for color matching step as in Figure 2.6. Pixels which belong to the ellipse mask are reserved while others are ignored. The reserved pixels were used to compute the color histogram. This color histogram is then compared to the color histogram of the people which was learned before. However, we could see that the ellipse can approximate the body and the head of the people very accurately. Therefore, the color appearance matching step benefits from this due to the decrease of ambiguous pixels from the background.

Figure 2.6 The ellipse mask for color matching

We compare the accuracy of approximating process between the rectangle human model and our proposed model in Figure 2.7.

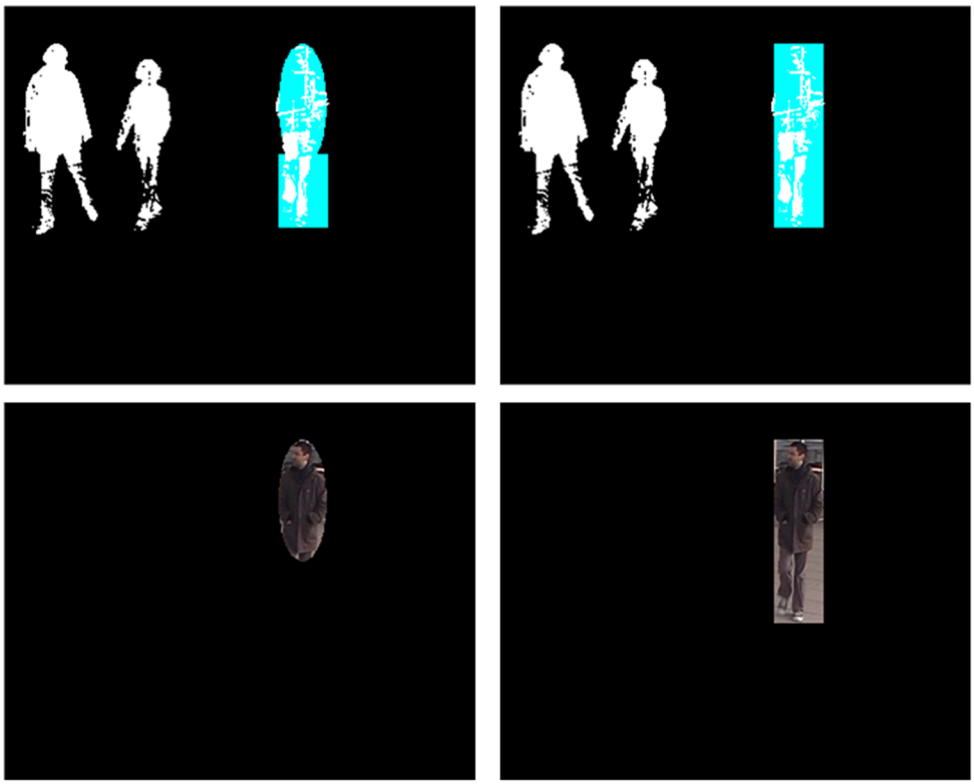

Figure 2.7 Comparing the accuracy between rectangle model and proposed model

In the above figure, the left column is the proposed masks which are used for background likelihood and color likelihood function. These two functions are the matching step comparing the similarity between the proposing hypothesis and the image which were captured from the cameras. This example is taken from one view. However, the right column of the figure is the masks when using the rectangle approximation. At the first row, we can see the proposed approximation is better than the rectangle approximation by enclosed almost the white pixels and ignored all the background pixels. On the other hand, the rectangle approximation still reserved a lot of background pixels, although it also quietly good in enclosing the foreground pixels. In addition, the row below shows that our model ignored almost of the ambiguous pixels from the background while the other still reserve a lot of them. The figure shows how our approximation model is more accurate than the rectangle model. We believe that our system can benefit this advantage in order to improve the accuracy of tracking process.

Moreover, note that our model is capable of have various sizes, which could help system to tracking almost of people from children to adults. The rotation angle of ellipse could be determined for extending purposes such as detecting the falling down people. However, this advantaged could be further research, and we did not

present it more detail in this dissertation.

2.3 Background subtraction

To generate binary foreground masks, we use a standard approach that maintains a mixture of Gaussians color model at each pixel [60], based on publicly available code due to Zivkovic [61] Adaptive background subtraction is a well-studied area. However, it is well known that the model does not handle sudden global illumination changes well, such as the sun coming out from behind clouds. For pre-recorded sequences, we use a forward-backward approach to compute foreground masks. Adaptive background subtraction is run forward in time from the first frame to last of the sequence, and a second process runs backward in time from the last frame to first. Assume a rapid change in illumination occurs at time t, with gradual illumination changes before and after that. The forward background subtraction pass will produce clean foreground masks for times less than t and then suffer degraded performance at time t before gradually recovering. Likewise, the backward pass will produce clean foreground masks for times greater than t, but suffer degraded results for a short period of time before t. Combining both the foreground and background masks with an AND operator tends to produce greatly improved masks both before, during and after the illumination change (see Figure 6). Although it may seem at first glance that this strategy is limited to batch processing, if a one to two second delay is acceptable then forward-backward processing methods can be performed on real-time camera streams by using a sliding temporal window approach [62].

2.4 CLEAR Metrics and Evaluation method

We compared our detection results against the POM+LP method, which is a multiple targets detection and tracking algorithm based on a probabilistic occupancy map and linear programming [12]. All of algorithms are evaluated based on the MODA (Multiple Object Detection Accuracy) and MODP (Multiple Object Detection Precision) metrics from the CLEAR evaluation framework [58], which we will now define. For a frame t, let G_i^t denote the annotated bounding box and D_i^t the detected box. The detection accuracy is counted as correct if the overlap ratio,

$$OL_i^t = \frac{|D_i^t \cap G_i^t|}{|D_i^t \cup G_i^t|} \quad (2.1)$$

is greater than some threshold τ . We systematically vary this threshold and compute the evaluation metrics at each threshold. Correct detections and false positives/negatives are determined by solving an assignment problem between the annotations and the detection output. Letting N_G^t be the total number of annotated objects, N_m^t the number of correct detections, m_t the number of false negatives, f_t the number of false positives, MODP measures the localization quality of the correct detections,

$$MODP_t = \frac{\sum_{i=1}^{N_m^t} OL_i^t}{N_m^t} \quad (2.2)$$

while MODA is a detection accuracy measure taking into account both false negatives and false positives,

$$MODA_t = 1 - \frac{m_t + f_t}{N_G^t} \quad (2.3)$$

For both metrics, larger values are better. We also evaluated the tracking accuracy based on Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP).

Multi-view data sets are intrinsically difficult to acquire with temporary setups, mainly due to the amount of necessary material and the problems inherent to the simultaneous manipulation of several recording devices. Besides, in most countries, filming people in public places is subject to very strict privacy protection laws. As a result, very few multi-view pedestrian data sets are currently publicly available. A notable exception is the recent PETS 2009 data set, which is made of pedestrian sequences filmed from seven different angles. In this thesis, we used two datasets for evaluation. The first is Terrace from the author of POM+LP and the second is Sparse sequence S2L1 from PETS 2009.

Cafeteria terrace sequences: Several sequences of more than three minutes were filmed on an outdoor cafeteria terrace in our campus. Four cameras were placed at the usual 2m high, at every corner of the area. In some of the sequences, up to nine people appear simultaneously in front of the cameras. On other sequences, tables and chairs have been placed in the center to simulate static obstacles. The session was shot early in the morning and the sun, low on the horizon, produces very long and sharp shadows on some videos.

Sparse sequence S2L1: We used four camera views, including one elevated, far field view (called View 1) and three low-elevation near field views with frequent, severe occlusions (Views 4, 5, 6, and 8). We first compute ground truth annotations by hand from four camera views, beginning by manually clicking on the head of each person in each view. This data set is very challenging for several reasons. First, lighting conditions are very poor, representative of what can be expected in a real-world surveillance situation.

Most images are under-exposed, except near the exits where they often are saturated. Second, the area covered by the system is large, which means that people can get very small when reaching the far end, making their precise localization challenging.

Chapter 3

Multi-view Pedestrian Tracking

In this chapter, we present in detail our detecting and tracking system. The system is composed by varying parts, from the background subtraction step to the hypothesis proposing step.

- The function of each part and the operation of the system are provided in section 3.1. In addition, we also designed the human model which approximates the appearance of pedestrians and being used for color matching.
- In section 3.2, we formulated our tracking algorithm by using the modified Bayesian model.
- In section 3.3, the fusion of the foreground mask in all view made the background likelihood function.
- Moreover, the step of learning and updating color appearance in section 3.4 shows its advantages to overcome the problem of identities switching.
- In order to smooth the trajectories and reduce the searching state space, some independent assumptions were composed in section 3.5, for example, the dynamic model, mutual exclusion, or target persistence.
- However, the marginal conditional probability could be intractable because of the complexity of the huge state space. Therefore, we use reversible jump Markov Chain Mote Carlo to battle this problem in section 3.6.

3.1 Overview of the system

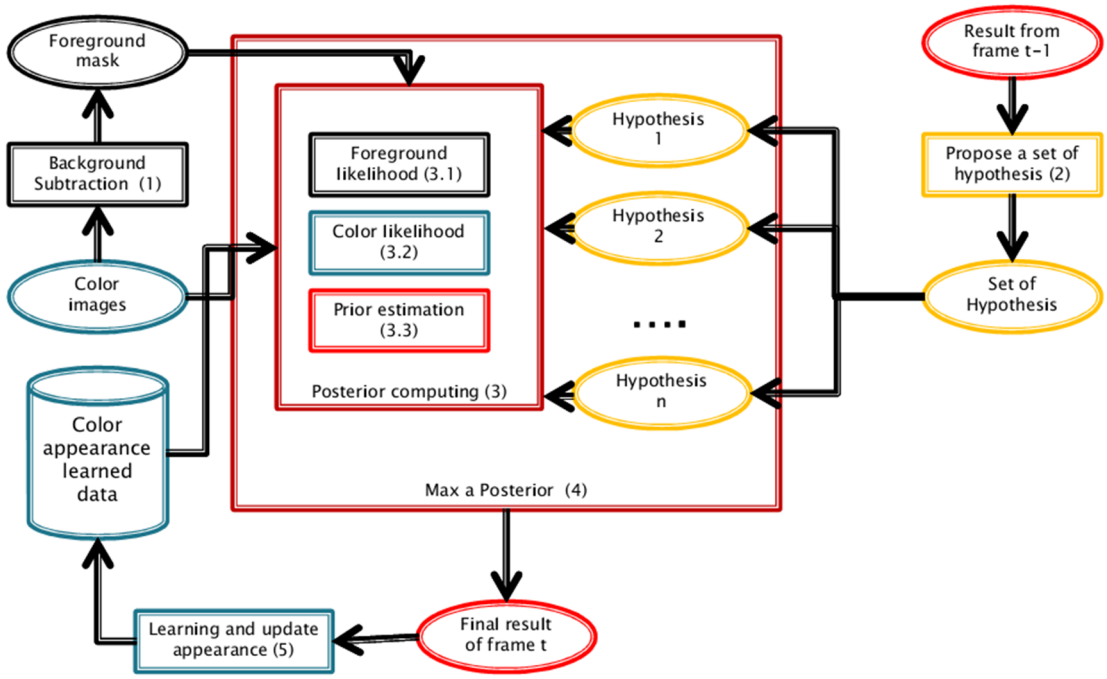

Figure 3.1 The tracking system composed by varying parts

The system illustrated in Figure 3.1 is composed by varying parts. In this figure, we refer to the operation within the system in one frame at time t. The inputs of the system are the color images observed from all camera views and the tracking result from the previous frame t–1. The result of pedestrian tracking is the accurate location of people in the observing ground where we could draw a rectangle to enclose the people in all views as in Figure 1.1. The system has to go through four steps before releasing the result by maximizing the marginal conditional probabilities. The first step is doing background subtraction to create the foreground mask from the color images of all views. The second step is proposing a set of hypotheses from the result of the previous frame. The third step is estimating the marginal conditional probability of each hypothesis in the above set. This step includes three sub steps, which are the prior estimation, the background likelihood estimation and the color likelihood estimation. Finally, based on the operation of third step, we find the max value of the marginal conditional probabilities in fourth step, which means the final tracking result for the frame at time t. This tracking result will be used as the input for the next loop at frame t + 1. However, after

estimating the final state, the system automatically learns and updates the color appearance of people in fifth step. These five steps are repeated frame by frame. The function of each step is described below.

- **Step 1 – Background Subtraction:** As mentioned in chapter 2, we used Zivkovic [61] Adaptive background subtraction methods to get the binary foreground mask. The advantage of the foreground mask is it can easily separate the occurrence of people in the scene. In Figure 3.2, we present the foreground masks of all four views which are computed by background subtraction step. The black pixels have value of zero belongs to the background of the scene while the value of one white pixels belongs to objects. However, the background subtraction method could raise much ambiguous information because of the clutter or the light illumination changing as the example in figure. The Figure 3.3 shows that some background pixels are still in the foreground mask while the real foreground pixels are ignored. The clutter occurs when the color appearance of the object has the same distribution with the color of the background. On the other hand, the temporary changing in light illumination created many noisy to the foreground mask. Improving the accuracy of background subtraction step is not included in our research. However, that our system could overcome this problem of noisy in background subtraction methods very well by using the dynamic model and the people persistent model.

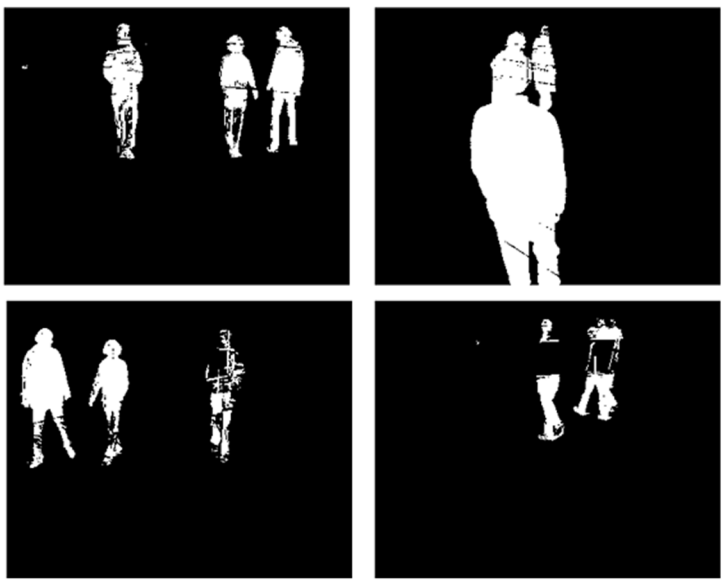

Figure 3.2 The foreground mask by Background Subtraction

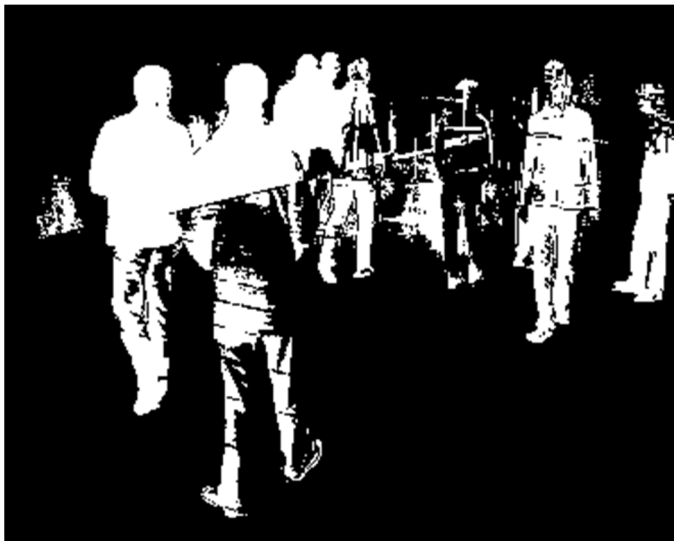

Figure 3.3 noisy from background subtraction

- **Step 2 – Proposing a set of hypotheses:** Within the original Bayesian formulation, operation of the system each time has to depend on the result of all previous frames. However, this approach requires a mass of computational operation due to their huge state space. This is also the problem of such Bayesian based methods like Multiple Hypotheses Tracking or Particle Filter. Nevertheless, now we can reduce the state space by using the modified of Bayesian model. The step of proposing a set of hypotheses in one frame is only depended on the result of the immediately previous frame. The example is presented in Figure 3.4. The left column is the result of a previous frame, while the right column proposes three new hypotheses for the current frame. We predict the location of people in the current frame around the previous location. For example, if the previous location of people in the grid is (x_0 , y_0), the predicted location will be (x_0+Δx , y_0+Δy), where Δx and Δy have the value belongs to [-2;2], However, to prevent the problem of lost tracking, sometimes we will predict of the people further than the normal neighbor around the previous location, for example (x_0+4 , y_0+4) This step will be described clearly in the update step of reversible jump MCMC step which approximate the marginal conditional probabilities.

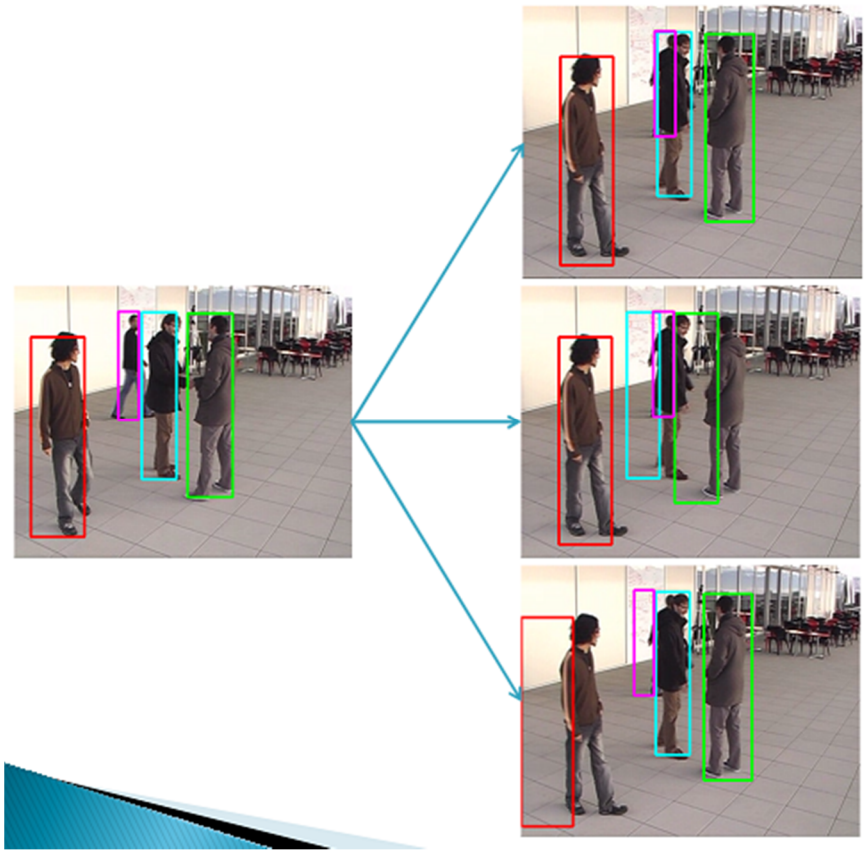

Figure 3.4 An example of proposing a set of hypothesis

- **Step 3 – Estimating the marginal conditional probability for each hypothesis:** We compute the probability for each predicted hypothesis by using inputs from color images and foreground masks of all views. As presenting above, this step has three sub steps.
  - Estimating the background likelihood: this sub step uses the information from the foreground mask to estimate the probability of the hypothesis. This step will be described in section 3.3.
  - Estimating the color likelihood: Color image and the database which stored the color of each tracked person will be the input of this step. Section 3.4 will show how it could make advantage in compare with the methods which used the foreground mask only.
  - Prior estimation: in reverse to 2 above sub-steps, this sub-steps is the estimation which based on the tendency of people motion. We have some independent assumption, which created advantages while do not affect too much on the generality of the system. Section 3.5 will present how the motion tendency could make the system work more effectively.

- **Step 4 – Max a Posterior:** After having probabilities of a set of hypotheses, we have already to release the result of people location by finding the max value. However, the system could be intractable when the state space is too large. For example, if there are six people in the scene, and each person has 25 neighbor locations, we have full state space at least 25^6 cases, which could not be directly estimated. Because of this problem, the reversible jump MCMC helps to overcome this problem by efficiently approximate the probability distribution of the searching state space. RJMCMC will be explained in section 3.6.
- **Step 5 – Updating and learning color appearance:** The result of tracking is ready for the fourth step. However, because the system has to learn or update the appearance of each person, which currently existed in the scene. If the people go to the scene first time, system learns their appearance. On the other hand, if they existed in previous time, the system can update their appearance. The reasons why we need the update step as well as the operations of color matching are described in section 3.4.

We had the overview of all functional for the system. However, the operation of all systems requires the clearer description of each function and the knowledge of the formulation. The remaining sections of this chapter will present all about the advantages of our proposed system.

3.2 Modified Bayesian based Formulation

In this section, we present the multi-view pedestrian tracking algorithm. We need to estimate the accurate location and the number of pedestrians in the scene given the color images, and the foreground masks obtained from the multiple cameras. At time t, the estimation is defined as $X^{*(t)}$ where

$$X^{*(t)} = \{ x_1^{(t)}, \ x_2^{(t)}, \ldots \ x_n^{(t)} \}, \text{ n} > 0 \qquad (3.1)$$

$X^{*(t)}$ includes the number of human objects n and their parameters $x_i^{(t)}$. Each i^{th} element $x_i^{(t)}$ represent the attribute of each tracked people which include their accurate location c_i in the tracking area j and their height h_i as

$$x_i^{(t)} = \{\ c_i,\ h_i\ \}$$

(3.2)

The location variable $c_i \in$ W specifies a 2D spatial coordinate position of people in the grid which was presented in chapter 2. In addition, despite fixing the generative human model at the constant number, our approach can determine the height of each target. This advantage makes the system become more flexible to the varying height of people, and therefore, increase the precision of the algorithm.

We formulate the tracking problem as computing the maximum a posterior (MAP) estimation.

$$X^{*(t)} = argmax_{X^{(t)} \in \omega}\ \mathrm{P}\ (X^{(t)} \mid I^{(t)}, X^{*(t-1)})$$

(3.3)

where $X^{(t)}$ is the estimation at time t proposed from each hypothesis and $I^{(t)}$ is the color images which captured from all camera views. The hypothesis $X^{(t)}$ is an element of the set of all hypotheses ω which we call the searching state space. On the other hand, $X^{*(t-1)}$ is the tracking result from the previous time t–1. From the result of $X^{*(t-1)}$ and given images from all views $I^{(t)}$, we need to estimate the posterior probability $\mathrm{P}\ (X^{(t)} \mid I^{(t)}, X^{*(t-1)})$ of each hypothesis $X^{(t)}$ in the searching state space ω. The hypothesis which has the max value of posterior probability will become the final tracking result $X^{*(t)}$.

The algorithm only takes into account images acquired at the same time by the multiple cameras. Its basic ingredient is a generative human model as in chapter 2, which represents humans as models of mixing rectangle and ellipse, and is used to create synthetic ideal images we would observe if people were at given locations. Under the images of a hypothesis, the posterior probability is estimated following to the modified Bayesian rule. The posterior probability of each hypothesis is decomposed into a likelihood term and a prior term.

$$\mathrm{P}\ (X^{(t)} \mid I^{(t)}, X^{*(t-1)}) = \mathrm{P}\ (X^{(t)} \mid X^{*(t-1)}) * \mathrm{P}\ (I^{(t)} \mid X^{(t)})$$

(3.4)

On the right term of the above equation, $\mathrm{P}\ (I^{(t)} \mid X^{(t)})$ is the likelihood probability, and $\mathrm{P}\ (X^{(t)} \mid X^{*(t-1)})$ is the prior probability. The prior probability $\mathrm{P}\ (X^{(t)} \mid X^{*(t-1)})$ is the probability that the new hypothesis $X^{(t)}$ depends on the previous tracking state $X^{*(t-1)}$. In our approach, the prior probability is the summarization of three physically motivated constraints and the ghost detection function, which will be presented more detail in section 3.5. On the other hand, the likelihood probability P

$(I^{(t)} \mid X^{(t)})$ is the probability that shows how the new hypothesis $X^{(t)}$ looks like to the image evidence $I^{(t)}$. Our likelihood probability is decomposed into the summarization of the background likelihood and the color likelihood.

$$\mathrm{P}\,(I^{(t)} \mid X^{(t)}) = \alpha * \mathrm{P}(BI^{(t)} \mid X^{(t)}) + \beta * \mathrm{P}(CI^{(t)} \mid X^{(t)}, color\_database) \quad (3.5)$$

In the right term of the above equation, $\mathrm{P}\,(BI^{(t)} \mid X^{(t)})$ is the background likelihood and $\mathrm{P}(CI^{(t)} \mid X^{(t)}, color\_database)$ is the color likelihood. The background likelihood $\mathrm{P}\,(BI^{(t)} \mid X^{(t)})$ shows how the foreground mask $BI^{(t)}$ looks like the proposed hypothesis which will be presented more detail in section 3.3. On the other hand, the color likelihood as in section 3.4 represents the probability of similarity between the color images $CI^{(t)}$ and the new hypothesis where the $color\_database$ term is embedded. The term of $color\_database$ here is the color appearance of each object existed in the tracking area. This database is automatically and realtime learned and updated when the people moved into the observing area. The two other parameters are α and β which are two constant numbers. These parameters represent the mutual contribution between two kinds of likelihood function. If we want the color likelihood takes more weight in the likelihood term, we need to set $\beta > \alpha$ which means that the contribution of color likelihood is larger than background likelihood.

We could see the different in Equation (3.4) in compared to the original Bayesian tracking problem. This modified Bayesian formulation does not have to predict step and also update step. It simply combined the detecting step and the tracking step into one unique Equation (3.4). However, the predict step is now responded to the Reversible Jump MCMC which is used to approximate the posterior distribution in the searching state space ω.

In addition, the Equation (3.4) shows that it does not need to combine all tracking results for a long time such as other original Bayesian based algorithms such as Kalman Filter or Particle Filter. On the other hand, the modified Bayesian posterior computation is only using t j he tracking result from one immediately previous frame. This advantage could help to reduce the searching state space which is the problem of previous approaches. The notation for all variables in the Bayesian formulation is presented in Table 3.1.

Table 3.1 The notation for variable in our modified Bayesian formulation

$X^{*(t)}$	The state of all objects after tracking process at time t
$x_i^{(t)}$	The state of the ith object
c_i	The grid location of ith object
h_i	The height of ith object
$I^{(t)}$	The color images from all camera views at time t
$BI^{(t)}$	The foreground mask from background subtraction at time t
$CI^{(t)}$	The color images from all camera views at time t
α	The factor of background likelihood
β	The factor of color likelihood
$color\_database$	The color of the human which is learned automatically

In the next section, we will present the background likelihood which is the joint likelihood based on the information of the foreground mask.

3.3 Background likelihood

The background likelihood term P ($BI^{(t)} \mid X^{(t)}$) in equation () shows how the new proposed state $X^{(t)}$ is similar to the image evidence from the foreground mask. The foreground mask here is the binary image which the pixels of zero belong to the background while other having value of one belong the object.

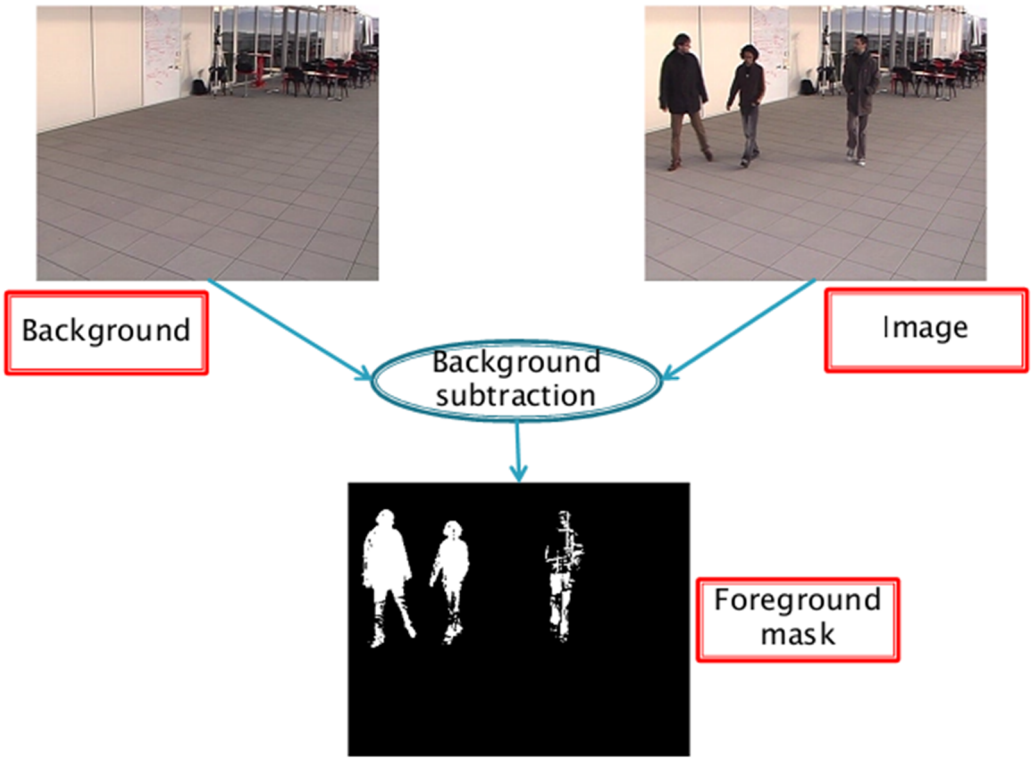

Figure 3.5 The background subtraction

For easily understanding, we refer to the Figure 3.5 which shows the inputs and the result of the background subtraction step. Firstly, we get the background image when there is no person in the scene. In tracking process, we get the color image from the camera. This color image is subtracted to the background image. The output of background subtraction step is the foreground mask. We can see the white pixels which have value of one belong to the objects. However, some white pixels are closely disappeared because of the similarity between color of the objects and color of the background. This problem is frequently called "clutter" Some present approaches have tackled in this problem due to the wholly depending on the foreground mask and ignore other evidences such as the motivated constrains as well as the color information. However, our approach helps to overcome by using the prior probability model as well as color likelihood probability.

The term of background likelihood is computing the similarity between the proposed hypothesis and the foreground mask from the background subtraction. We predict the state of all objects based on the information of previous tracking results. However, we now recall the definition of the state of each object $X^{(t)}$ which was presented in section 3.2.

$$X^{(t)} = \{ x_1^{(t)} , x_2^{(t)}, \ldots. x_n^{(t)} \}, \text{n} > 0$$

The parameter n is the number of objects which existed in the scene at time t. Each i[th] object $x_i^{(t)}$ represents the attribute of each person, which include their accurate location c_i in the tracking area and their height h_i as $x_i^{(t)} = \{ c_i, h_i \}$. By using this state of the hypothesis, we create the synthesis image which transforms information of the state to the image evidence. The image evidence for background likelihood is the gray-scale image where each pixel has a value ranges from 0 to 255. To create the synthesis image, we firstly project the state of each object, the grid position as well as the height, into all view where the people can be seen by the camera. The generative human model which was described in chapter 2 is used for this projection. The example of this projection for the appearance of one object in all four views is shown clearly in Figure 3.6. In the figure, the cyan pixels which enclose the white pixels are pixels which were projected by the state of an object. We could see that the cyan area perfectly approximate the area of white pixels which belong to the object. This could be seen with naked eyes evidence shows that our proposed hypothesis could be the accurate state of the object.

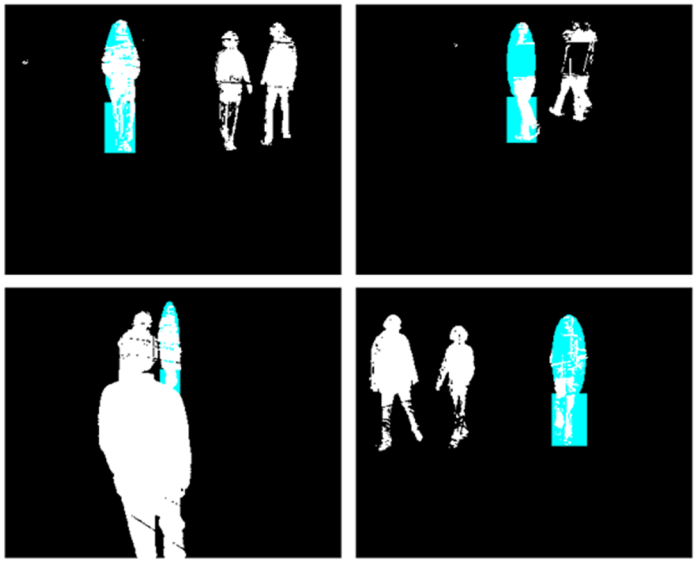

Figure 3.6 The projection of human model approximated the foreground mask

However, since multiple humans may occlude each other, the image likelihood cannot be decomposed into the product of image likelihoods individual human hypotheses. Given a state, we have to compute joint likelihood based on the formation of the foreground by fusing information about all objects in the scene. Now, the synthesis images are composed by projecting all objects. In Figure 3.8, we show the result of synthesis images which were projected all objects. There are three persons in the figure, the most left images are the foreground masks from

background subtraction. The middle images are synthesis images which were created by projecting all three persons to the camera views. The most right images show how the synthesis images similar to the foreground masks by overlapping the synthetic images to the foreground masks. The synthesis images are done for all four views.

As described above, synthesis images are gray-scale images, which mean each pixel in the image has been varying value ranges from 0 to 255. The reason for choosing this kind of synthesis image is that it could estimate the occlusion between multiple human. If we only used the binary image such as the foreground mask, the occlusion problem could not be solved completely because of ambiguous information. The pixel which has the position that could belong to two people actually has the larger weight than another, which just belongs to one people. Way of modeling hypothesis helps us to overcome the problem of occlusion. The example is described in Figure 3.7, where the larger weight pixel is presented by the darker cyan pixels. In this figure, the brightest cyan pixels belong to the only one person, while the darkest pixels were created by the overlapping of three persons. As the larger weight the pixel, the probability when the pixel at the same position in the foreground mask is the white pixel is also bigger than other.

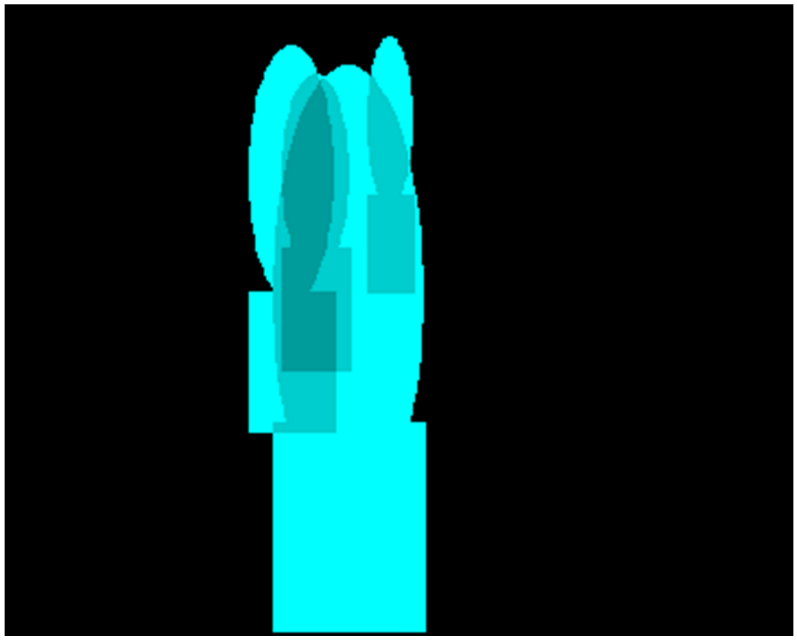

Figure 3.7 Synthesis image as gray-scale image which having different value at each pixel

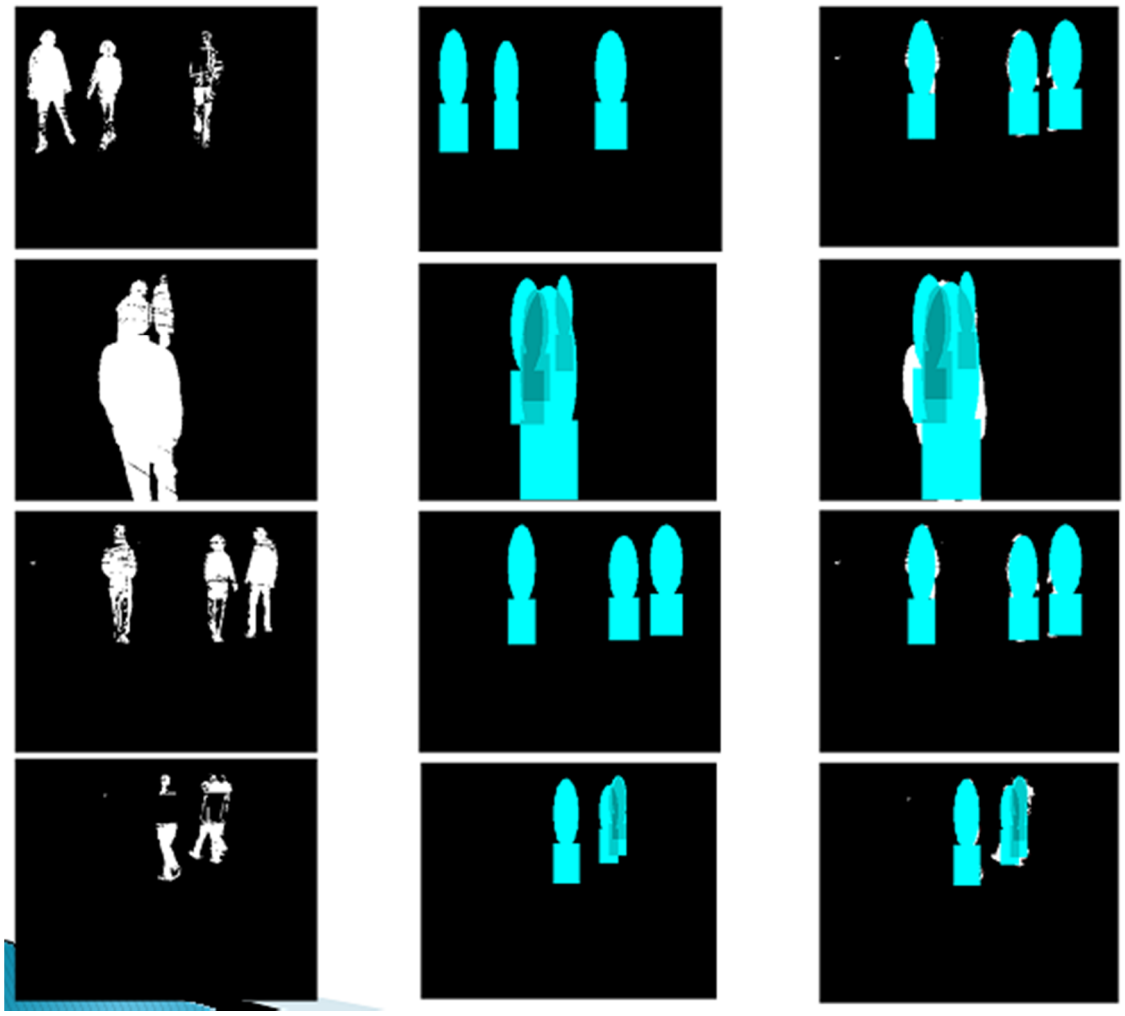

Figure 3.8 Synthesis images for all views

After having the synthesis images of all views, given the foreground masks, we compute the background likelihood probability. For fusing all camera views, the general background likelihood probability is computed by summarizing the background likelihood probability of all camera views. We have the equation:

$$\mathrm{P}\left(BI^{(t)} \mid X^{(t)}\right) = \sum_{v=1}^{V} e^{-\rho\left(BI_v^{(t)}, SBI_v^{(t)}\right)} \qquad (3.6)$$

where V is the number of camera views and $\rho(BI_v^{(t)}, SBI_v^{(t)})$ is the normalized pseudo-distance to account for the similarity between the foreground mask $BI_v^{(t)}$ and the synthesis foreground image $SBI_v^{(t)}$ of view v. For each view, the normalized pseudo-distance between the foreground mask BI and the synthesis foreground image SBI is computed by the equation:

$$\rho(BI, SBI) = \frac{\sum_{i=1,j=1}^{Width,Height} p_{ij}^{BI} * (1 - p_{ij}^{SBI}) + p_{ij}^{SBI} * (1 - p_{ij}^{BI})}{\sum_{i=1,j=1}^{Width,Height} p_{ij}^{SBI}} \quad (3.7)$$

Where p_{ij}^{BI} and p_{ij}^{SBI} are the value of the pixel at the position (i, j) in the foreground mask BI and the synthesis foreground image SBI consequently. The numerator term is the computation of the distance or the difference between two images. This distance is computed by summarizing the difference of value between all pixels in two images. On the other hand, the denominator is called as the normalized factor which scaling the normalized distance. The normalized factor is the summarization of pixel value in the synthesis foreground image.

Given the background likelihood, we can estimate the probability that the new hypothesis meets the image evidence. Actually, that is the probability that the synthesis images similar to the foreground masks.

However, there are also some problems that make the tracking process lose the objects by using only the background likelihood. As fore-mentioned in section 3.1, background subtraction step, which creates the foreground masks have to suffer many noises from clutter and illumination change. Therefore, the foreground mask still includes a lot of ambiguous pixels, which affect directly to the accurate of background likelihood. For example, suppose that our proposed hypothesis is totally accurate in estimating the grid location of the person. Then, in the foreground mask, pixels which positions are enclosed in the human model should have the value one. However, because of noise, those pixels are zero, and it increases the difference between the synthesis images and the foreground masks. The background likelihood probability is also decreased regardless the perfect estimation of our hypothesis. To overcome this problem, we get the new idea when embedded the color appearance in cooperation to the background likelihood. The foreground masks could have ambiguous pixels from noise but the color images.

Table 3.2 Notation for background likelihood computation

$BI_v^{(t)}$	The foreground mask from background subtraction for view v
$SBI_v^{(t)}$	The synthesis image for the hypothesis at view v
$\rho(A, B)$	the normalized distance between two foreground images
p_{ij}^{I}	The value of the image I at pixel (i, j)

3.4 Color likelihood – color appearance learning and updating

In order to improve the tracking accuracy and overcome some problems of those approaches using only the foreground masks, we present here the usage of color in the research of multi-views multi-targets tracking. As described before in chapter 2, the ellipse is used for approximating the human torso. As we know, when persons moving, the torso still has the rough shape while the others such as legs or hands have not. We can see visually via the below figure.

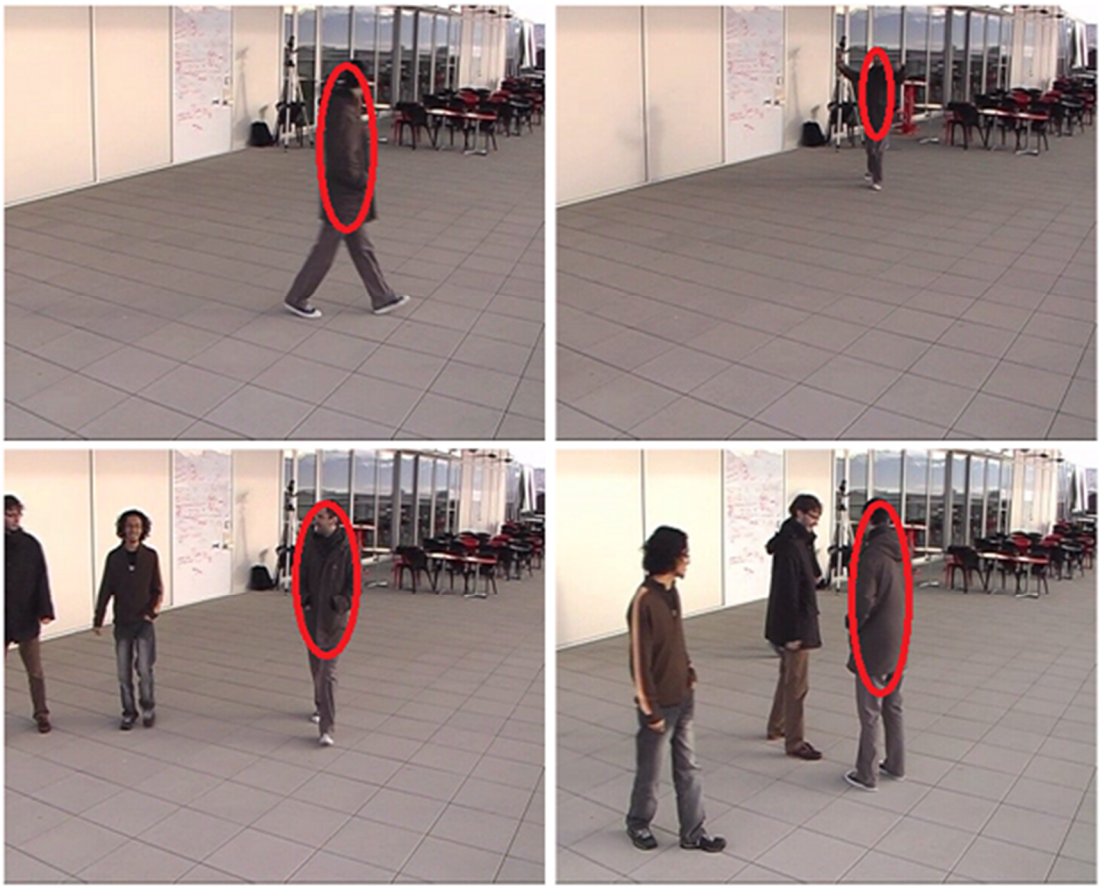

Figure 3.9 Ellipse shape for color matching

The above figure shows four kinds of the shape of the person. He is walking from left to right in the top left image, while in the top right, he raised his hand. In the bottom left image, he is walking from far to near, while in the bottom right, he stops walking and talks to his friend. In all four cases, we realized that there is no rough shape that could approximate all the human body without getting the ambiguous

pixels from the background. However, in our approach, we only used the ellipse to approximate the human torso which does almost not change the shape regardless of the human motivated situations. Approximating the torso by using the ellipse could help to reduce a lot of ambiguous pixels, and that is the reason why we choose the generative human model as the mixing between the rectangle and the ellipse. The whole human model was used in the background likelihood computing, but in color likelihood, we only used the ellipse part of the model to calculating the color histogram as well as matching.

The ellipse itself can be used for color histogram calculating, however, our approach divided the whole torso ellipse into seven different parts. We realized that the spatial information is available by using this division. Seven parts of the ellipse are presented in Figure 3.10. The first part is the whole torso. Then, the torso is divided into two smaller parts. The center of the torso is the second part, while the third part is the side where the torso touches to the background. The four other parts in the most right ellipse are used in case of a part of the torso is occluded by other people. The example is presented in the figure 3.11 where half of the person is occluded by another. Therefore, there are only the fifth and the sixth part being available for color matching.

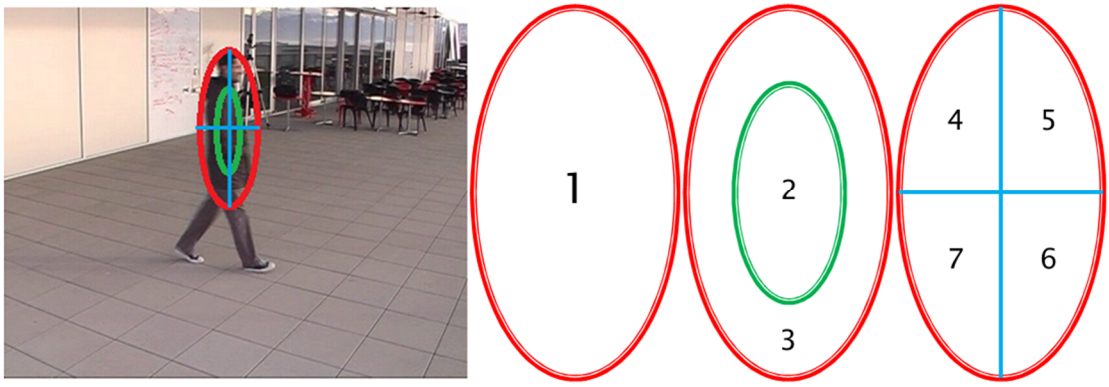

Figure 3.10 Seven parts of ellipse model

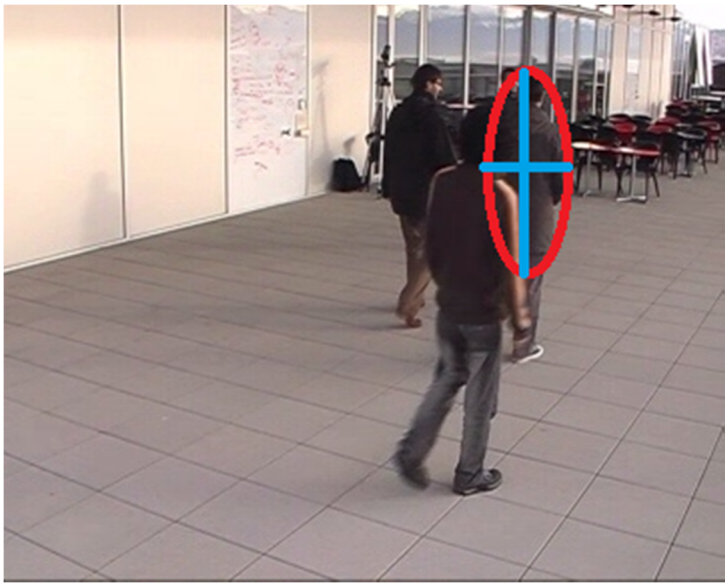

Figure 3.11 Visible and Invisible parts of ellipse

Recalling the section 3.1 where the whole system process was step by step defined. Particularly, the color likelihood receives three inputs, which are the proposed hypotheses, the color images from all views and the information from the color database. The color likelihood function returns the probability how the new hypothesis matches the color images' evidence given the color database. After each frame time, the color appearance of each person is automatically learned and updated for the next frame usage. However, this color matching step has been done separately each view before being summarized to contribute the final color likelihood probability following the below formulation:

$$\mathrm{P}(CI^{(t)} \mid X^{(t)}, color\_database) = \sum_{v=1}^{V} e^{-\sigma(CI_v^{(t)} \mid X^{(t)},\ color\_database_v^{(t)})} \quad (3.8)$$

On the right term, the function $\sigma(CI_v^{(t)} \mid X^{(t)},\ color\_database_v^{(t)})$ return the match probability for view v by combining three inputs the color image $CI_v^{(t)}$, the new proposed hypothesis $X^{(t)}$ and the information of color database of view v at time t.

The process of matching color for each view is presented in the below figure.

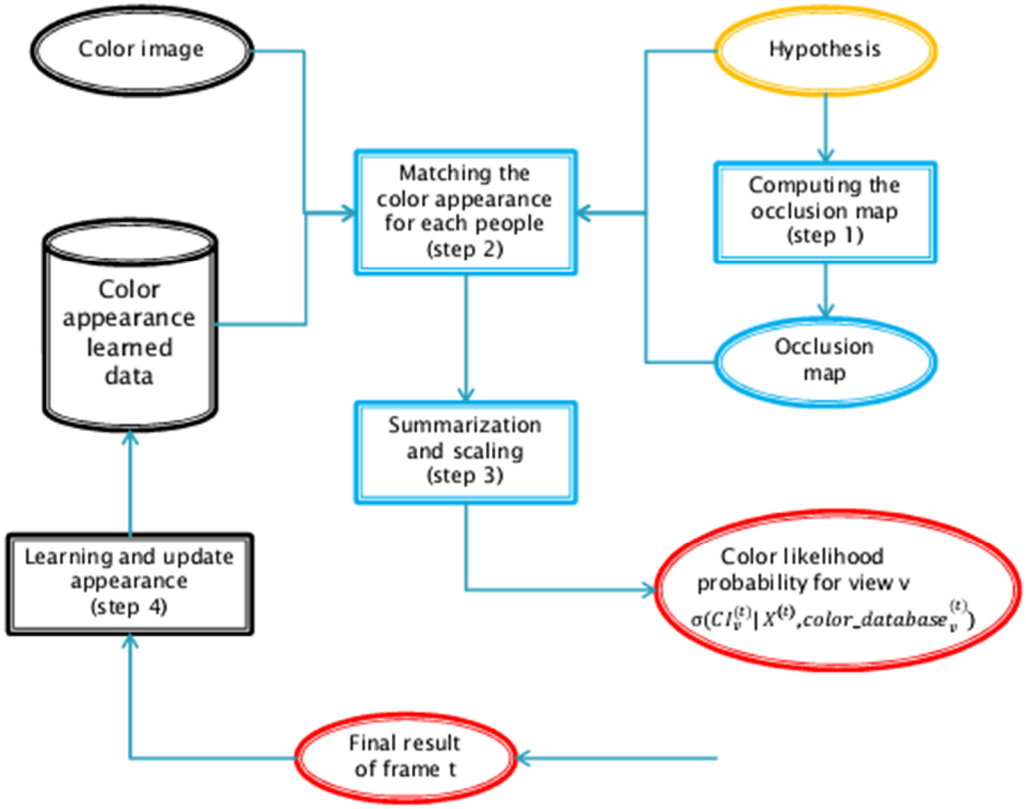

Figure 3.12 The process of matching color for each view

There are three steps before releasing the color likelihood probability for each view, while the fourth step learns and updates the color appearance for all people in the view given the final tracking result. First of all, we compute the occlusion map of the hypothesis via step 1 by using the camera calibration model which was introduced in chapter 2. The occlusion map is the information about mutual occlusion between all targets in the view. It points out that in that view which target is closer to the camera than other. It also means that the color pixels in the color image belong to the ellipse area of the closer target is not used to calculate the color histogram for other targets. The occlusion map, moreover, gives the information about which persons exist in the camera views. Those invisible to the camera view will return the probability of 0. Step 2 of the process receives the hypothesis state, the occlusion map of hypothesis, color image from camera view and the information from the color database. For each person who is available in the camera view and given the hypothesis, we project the ellipse approximate their torso as presented above. In the ellipse, for each part which is not occluded by other persons, we compute the color histogram for pixels those lied in the area of the part. The histogram of each part is then matched to the corresponding part of the person

from the color database. The color matching step for each person available in the camera view is done by the below equation.

$$\tau\left(\mathrm{CI}\middle|x^{(t)}\right) = \sum_{p=1}^{7} \delta_p BH\left(image\_his_p^{(t)}, database\_his_p^{(t)}\right) \quad (3.9)$$

There are seven parts in the ellipse, but we only compute those parts which are not occluded by other persons. Therefore, the $BH(image\_his_p^{(t)}, database\_his_p^{(t)})$ is available if and even if the histogram of the part in both the color image and the database are available. The factor δ_p is the factor which determines the weight of each part in contributing to the summarization. $BH(image\_his_p^{(t)}, database\_his_p^{(t)})$ is the Bhattachayya distance between two histogram models, the histogram of the hypothesis and the histogram from the database.

After having the color matching probability for each person, the summarization and scaling for the set of all targets is done in step 3. The formulation for this step is below:

$$\sigma(CI_v^{(t)} | X^{(t)}, \ color\_database_v^{(t)}) = \frac{\sum_{i=1}^{n} \tau(CI_v^{(t)} | x_i^{(t)}) * visible_i}{\sum_{i=1}^{n} visible_i} \quad (3.10)$$

The factor $visible_i$ is a binary number which has value of one if the person is visible in the camera view and value of zero in the reverse case. The denominator is the number of all people who are visible in the camera view, and this is also the scaling factor to normalizing the probability of color likelihood.

At step 4, after receiving the final tracking result from the system, the learning and updating step is executed. If the person is first time detected in the scene, we learn it. On the other hand, if he or she was detected in previous frames, we need to update the color appearance. Similar to the matching step, the learning and updating step is also processed by each view separately. This process is different to those approaches which are trying to reconstruct the human 3D shape by using complete calibration cameras. These approaches tried to fusing all color information of all cameras to create the 3D voxel images which could be used to reconstruct the visual hull. However, this process requires a lot of computation due to the color corresponding step. Therefore, preventing the different color distribution on each view and reducing the weight of computation is the reason that we need to ignore the comprehensive 3D reconstruction. We could see the difference in color

distribution on each camera view in Figure 3.13. The position of each camera affected directly the illumination of the color distribution, which make the problem of difference.

Figure 3.13 The difference in color distribution between different views

Table 3.3 Notation for color likelihood

σ	The matching function between the color image and the color database given the hypothesis at each view
τ	Color matching for each person including 7 parts
δ_p	The weight factor for color matching of each part
$BH(his1, his2)$	Bhattachayya distance of two color histogram

3.5 Prior distribution

As mentioned in the top of the chapter, besides the process of estimating the likelihood probability, some independent assumptions are added to the system in order to increase the accuracy as well as smooth targets' trajectories.

Our assumptions are related to two physically motivated priors and the ghost detecting prior. Two physically motivated priors include the object dynamic model

and mutual exclusion model. Therefore, the prior probability is decomposed into the product of three terms as in the equation below.

$$\mathrm{P}(X^{(t)} \mid X^{*(t-1)}) = e^{-\mathrm{dyn}(X^{(t)}|X^{*(t-1)},X^{*(t-2)},\dots)} * e^{-\mathrm{exc}(X^{(t)})} * e^{-\mathrm{gho}(BI^{(t)}|X^{(t)})} \quad (3.11)$$

For the motion term $\mathrm{dyn}(X^{(t)}|X^{*(t-1)})$, we use a constant velocity model. The dynamic model can be interpreted as a kind of "intelligent smoothing", which is actually more intelligent than those approaches blindly connect the nodes of the trajectory curve. It also helps to prevent identity switches between crossing targets by using their favors straight paths. Note that the dynamic model has so far been a weak point of trackers based on ILP. These methods suffer from aliasing of the discrete location grid, and either had to discard the dynamic model altogether. However, the dynamic model has to use information about velocity more than one previous frame in order to get the favor paths of the human. Therefore, we process the dynamic model on the batch of fifteen frames, which is long enough for estimating the temporary trajectories. The dynamic model is formulated by the equation below:

$$\mathrm{dyn}(X^{(t)}|X^{*(t-1)}, X^{*(t-2)}, \dots) = \sum_{i=1}^{n} \sum_{f=t-15}^{t} \left| v_i^f - v_i^{f+1} \right|^2 \quad (3.12)$$

where i is the No. of object and f is the No. of frame. The dynamic probability is the summarization of the dynamic model of all targets in the batch of fifteen frames. v_i^f is current velocity vector of target i where $v_i^f = c_i^{f+1} - c_i^f$.

The most obvious physical constraint is that two objects cannot occupy the same space simultaneously. We include this constraint into the prior estimation by defining a mutual exclusion term:

$$\mathrm{exc}(X^{(t)}) = \sum_{i \neq j} \frac{close\_dist^2}{|c_i^t - c_j^t|^2} \quad (3.13)$$

with the scale factor $close\_dist$ which is set to 1 on the grid distance (equavilent to 12.5 cm) for people tracking. Configurations are penalized where two targets come too close together, and the value goes to infinity when the y share one identical

position. The term at the same time enforces unique data association (since each detection can only be assigned to one trajectory).

This formulation of collision avoidance takes into account the actual overlap of target volumes and can correctly handle two notoriously difficult problems of multi-targets tracking: on one hand, overlap between targets is checked at all times, even if both targets are occluded. On the other hand, if two targets would collide due to inaccurate tracking, the mutual exclusion model can push them apart just as much as needed.

The ghost detection term $\mathrm{gho}(BI^{(t)}|X^{(t)})$ is the extra penalty. Given the number of foreground pixels covered by the projection model of each human state, the ghost detection determines whether a person hypothesis is valid or not. The ghost penalty for each person $x_i^{(t)}$ is formulated by the below equation.

$$\mathrm{gho}(BI^{(t)}|\, x_i^{(t)}) = \begin{cases} 1 & \exists v, s.t.\ \frac{|SBI^v(x_i^{(t)}) \cap BI^v|}{|SBI^v(x_i^{(t)})|} \leq ghost\_const \\ 0 & otherwise \end{cases} \quad (3.14)$$

The ghost detection penalty for each person is computed by counting the number of foreground pixels in the foreground mask BI^v covered by the human hypothesis in the synthesis foreground image $SBI^v(x_i^{(t)})$. The $ghost\_const$ is the penalty constant which value is 0.1. The ghost detection for all targets is defined as the summarization of ghost penalty of each target.

$$\mathrm{gho}(BI^{(t)}|X^{(t)}) = \sum_{i=1}^{n} \mathrm{gho}(BI^{(t)}|\, x_i^{(t)}) \quad (3.15)$$

If existing one view which the number of foreground pixels in the foreground masks less than 10% the number of pixels covered by the projection from synthesis image, the human hypothesis should be invalid. It increases the ghost penalty, and therefore, decreases the probability of prior estimation as well.

However, by the prior estimation, the trajectory tracking process is smoothed by the knowledge about human motivation as well as the error in proposing hypothesis given the observation images. Therefore, it helps to increase the accuracy, and reduce the searching state space.

Table 3.4 Notation for prior distribution

dyn	The dynamic model
v_i^f	The velocity of ith object at frame f
exc	Mutual exclusion
$close\_dist$	The penalty distance between two objects
gho	Ghost detection
$SBI^v(x_i^{(t)})$	The synthesis image given the state of one object

3.6 Reversible Jump Markov Chain Monte Carlo (RJMCMC) approximating the posterior distribution.

To perform Bayesian inference of the best configuration of human state, we maximized the posterior distribution which is the product of the prior term (Eqn. 3.11) and the likelihood (Eqn. 3.5). Finding the mode of the resulting posterior then provides a MAP estimate over the configuration space. Although direct computation of the posterior distribution is intractable, we use reversible jump Markov Chain Monte Carlo (RJMCMC) to battle this problem. Recent developments of MCMC methods have led to major advances in the simulation and inference of approximating the complex distribution, enabling us to work on relatively large hypothesis patterns.

MCMC methods were originally developed to generate samples from complicated target distributions, such as our posterior distribution that has an intractable normalizing constant. The algorithm constructs a Markov Chain with the desired target distribution as its equilibrium distribution. RJMCMC [63] extends the classic algorithm to deal with variable dimension models and suits the crowd analysis problem well because the number of people is not known a priori and thus also needs to be estimated. In the RJMCMC framework, searching for the optimal configuration of a fixed number of people is equivalent to hypothesis testing, while determining the number of people in the scene is treated as a Bayesian model selection problem.

RJMCMC is an iterative sampling procedure that involves proposing local updates to a current configuration or a reversible jump between configurations of

differing dimensions, and then deciding stochastically whether or not to accept the new configuration based on the value of the acceptance ratio:

$$\alpha(\mathrm{X}, \mathrm{X}') = \min\left(1, \frac{\pi(X')Q(X;X')}{\pi(X)Q(X';X)} * J_{F_{|X|\to|X'|}}\right) \quad (3.16)$$

Where the target distribution π is the posterior distribution $\mathrm{P}(X^{(t)} \mid I^{(t)}, X^{*(t-1)})$ in equation (), X and X' are the current and the proposed configurations, Q (a, b) is the probability of proposing a transition from a to b, and J is the Jacobian determinant of a dimension matching function F [63], which for us simplifies to a constant value of one.

Starting with an initial state, RJMCMC proposes a new state X' from a proposal distribution $Q(X';X)$ that depends on the current state X. The proposed state is probabilistically accepted as the next state in the Markov Chain according to the acceptance ratio. The design of good proposal distributions is the most challenging part of the sampling algorithm. Proposals that only allow local perturbations may become trapped in local modes, leaving large portions of the solution space unexplored, whereas global adjustments have less chance to be accepted unless the target distribution is very smooth or tempered to be so.

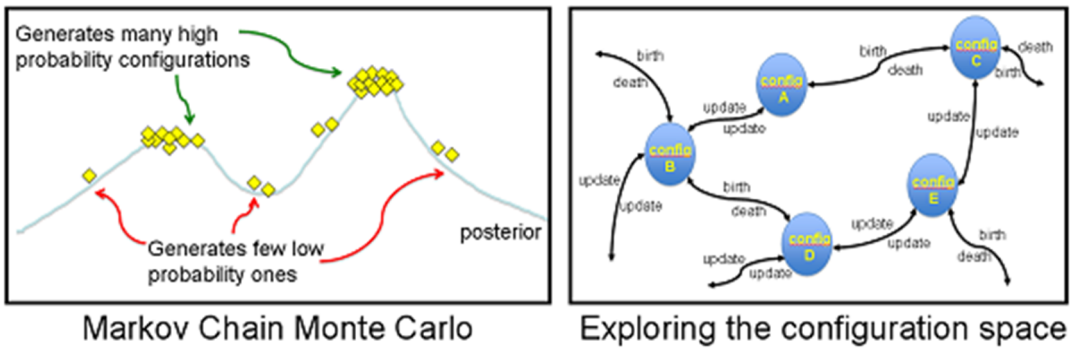

Figure 3.14 MCMC explores the configuration space

The RJMCMC sampler explores the space of configurations by proposing perturbations to a current configuration, as illustrated in Figure 3.14. We use a simple RJMCMC sampler composed of birth, death and update proposals [64]. Each of the proposals is described briefly below:

Birth/Death proposal: A birth proposal adds a 3D person to the current configuration, i.e $X' = X \cup x_b$. A simple birth strategy might place a person uniformly at random (u.a.r.) in the bounded region W. A death proposal removes a person x_i from the current configuration so that $X' = X/x_d$, e.g. choosing x_d u.a.r.

from X. Both proposals involve a dimension change from $|X|$ to $|X'|$. Instead of blindly adding a person, we use a more informative data-driven proposal [65]. We sample x_b's location according to the birth probability

$$P_b = \frac{P_b(l)}{\sum_{l \in W} P_b(l)} \tag{3.17}$$

where $P_b(l) = \sum_v \frac{|SBI^v(l) \cap BI^v|}{|SBI^v(l)|}$ is the fused occupancy likelihood of a particular location l, computed as the sum of the percentage of foreground pixels within its projected model in all views, and W is a bounded region of interest which is defined as entry/exit place. At the beginning of each frame time, the birth probability for the entrance is computed. Given this probability, the new targets are added into the current state. On the other hand, those existed targets which location is in the exit location, the death proposal proposes to remove the target out of the current state.

Update proposal: The update proposal preserves the dimension of the cur-rent configuration but perturbs its member's attributes (location and size) to gen-erate a new configuration. We use a random walk proposal that selects a person x_u u.a.r. from X, and either proposes a new spatial placement by sampling from a truncated normal distribution $\mathrm{N}(c'_u|c_u, \sigma)$ centered at the current location c_u, or proposes a new height by sampling from a truncated normal centered at the height of an average person, h=1.7m. Since the Gaussian distribution is symmetric, $Q(X'; X) = Q(X; X')$. Hence, the acceptance ratio simplifies to α(X, X') = min(1, $\frac{\pi(X')}{\pi(X)}$).

For all the experiments, we always start with an empty configuration as the initial state. The RJMCMC procedure is iterated about 100*number-of-object times, with the larger number of iterations being needed when there are more people in the scene. The move probabilities for birth, death and update proposals are set to be 0.15, 0.15 and 0.7, respectively.

Chapter 4

Evaluation

As described in chapter 2, we evaluate our algorithm on the PETS2009 benchmark dataset and the terrace dataset from the author of the POM+LP method. The camera calibration information provided with each dataset was used to generate the birth proposal map as well as the average back projection of foreground masks from all views.

We compared our approach with POM+LP method, which is the state of the art in multi-target pedestrian tracking. This method is a multi-target detection and tracking algorithm based on a probabilistic occupancy map and linear programming. We used the CLEAR metric MODP and MODA to evaluate the tracking precision. Our proposed method obtains superior results compared to POM+LP method, as will be shown through quantitative evaluation below.

The advantage of our approach is the color likelihood. In order to point out the better result, we also created another system, which used only the background likelihood estimation. Then, we compared POM+LP method to our two methods. The first of our method did not use the color likelihood while the other used it. By this comparison, we show how effectively the color appearance was used in our approach to improve the tracking precision.

4.1 Multiple Object Detection Precision (MODP)

The chart in Figure shows the MODP quantitative evaluation for two datasets. We could see that the state-of-the-art POM+LP method reached very high performance with over 80% of accuracy. Our method which did not use the color likelihood became weaker than POM+LP. The reason for this weakness is while POM+LP required the optimization for the batch of frames; we only used the tracking result from one previous frame. However, because of requiring the information from the longer tracking, POM+LP takes too many times for computation as well as could not be applied to real-time situation. On the other hand, our methods could be used in real-time because it only needs information from a very short interval. However, when we added more information, the color appearance, our method became better than POM+LP method. We can see it clearly

when dealing with some noise from background subtraction. While POM+LP got a lot of ambiguous results, our approach was not affected too much by using the color information.

The figure shows us more clearly about the accuracy between methods. In the terrace dataset, POM+LP got a very high result of over 81%, but in PET dataset, they have just got the result about 62% due to the noise of dataset. These noises came from the calibration error, and the lighting condition. They are also problems of our method. However, in both datasets, our approach with color likelihood still got better results than POM+LP.

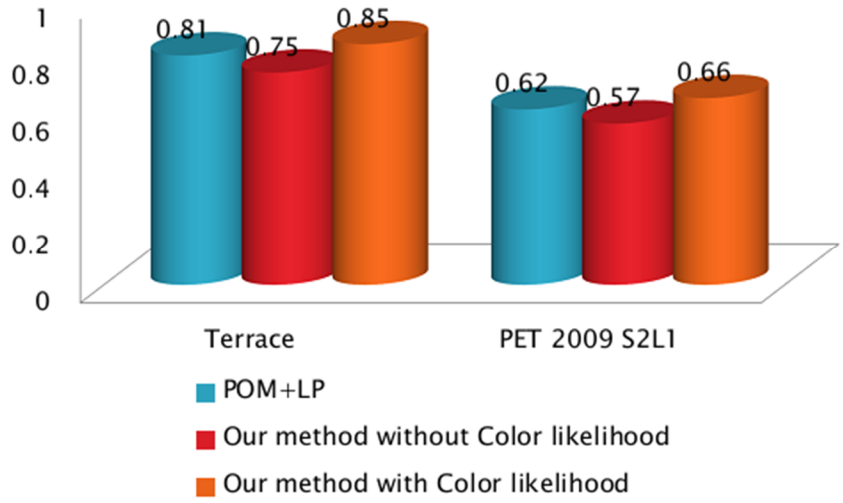

Figure 4.1 The MODP evaluation between POM+LP and our proposed methods

4.2 Multiple Object Detection Accuracy (MODA)

We also used MODA quantitative metric for evaluation. The comparing way and the result are quite similar to above MODP metric. Our method with color likelihood still got the better a result than POM+LP/ The result is shown in Figure 4.2.

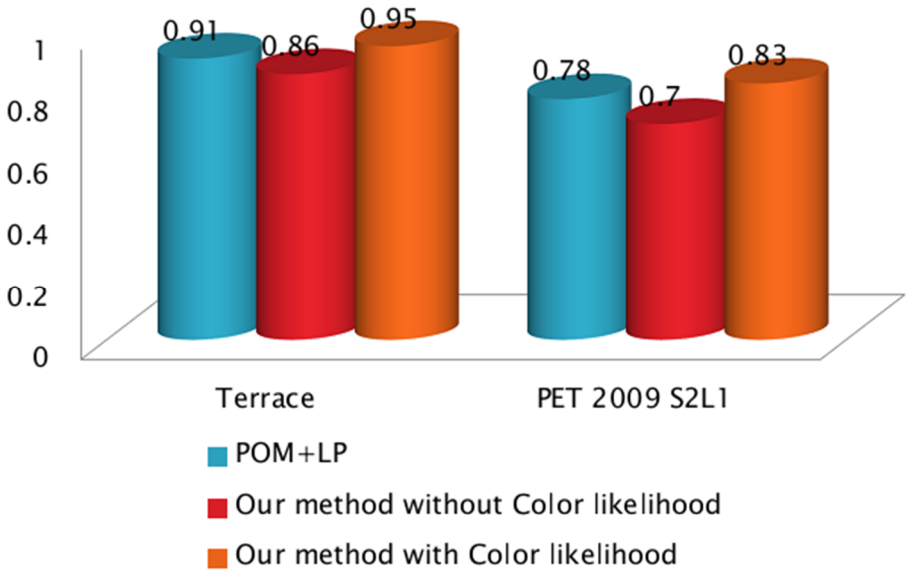

Figure 4.2 The MODA evaluation between POM+LP and our proposed methods

4.3 Determining α and β

Two parameters α and β in equation 3.5 accounted for the mutual contribution of two kinds of likelihood function, the background likelihood and the color likelihood. If we want the color likelihood takes more weight in the likelihood term, we need to set $\beta > \alpha$ which means that the contribution of color likelihood is larger than background likelihood. The ratio of the contribution between the color likelihood function and the background likelihood function is set as $\gamma = \frac{\beta}{\alpha}$. This is the only parameter that needs to be tuned in our approach. Empirically, we have found that a value of 1 gives the best results in almost all situations. This means the contribution of two likelihood functions should be equal. Figure (4.3) and (4.4) illustrated this fact experimentally, by plotting detection precision and accuracy as a function of γ. As can be seen, the optimal value is closed to 1 for experiments of both datasets.

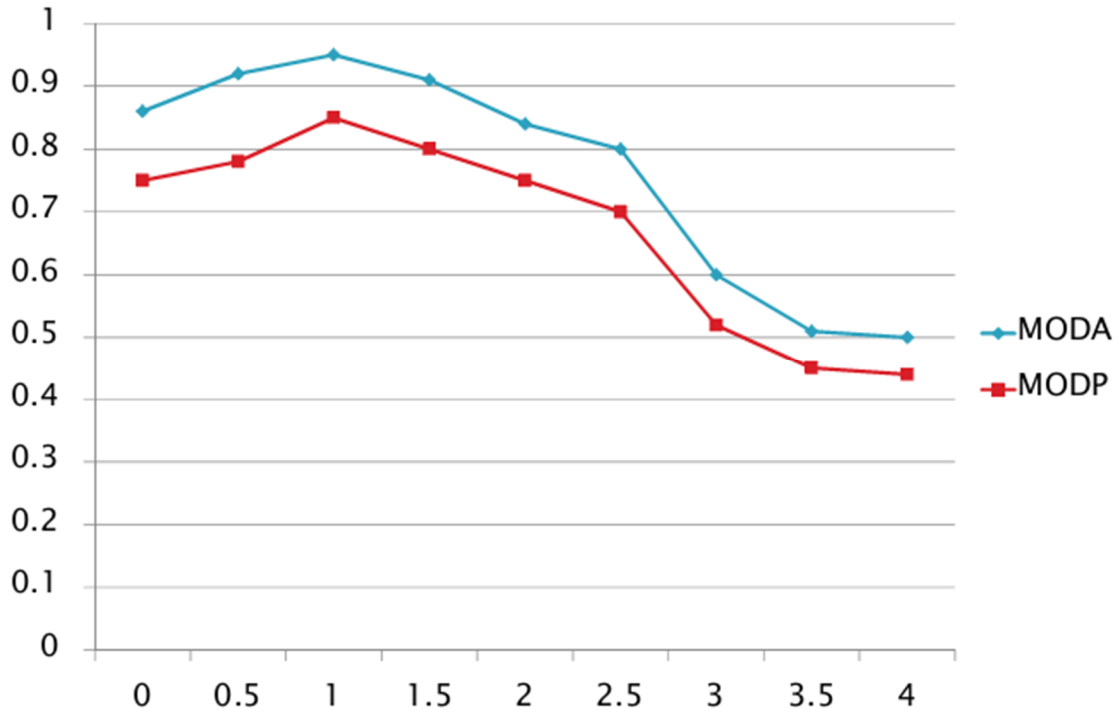

Figure 4.3 Illustration for changing the ratio in Terrace dataset

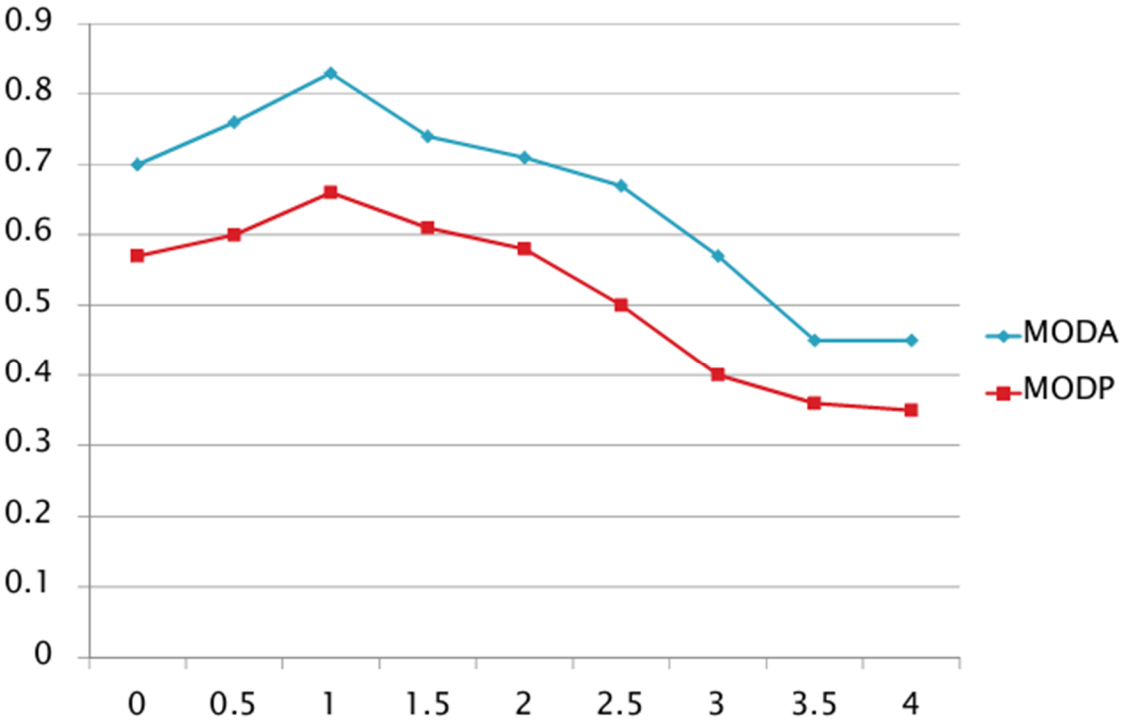

Figure 4.4 Illustration for changing the ratio in Pet 2009 dataset

4.4 Run Time

The last remaining aspect of the evaluation is the speed of our algorithm. To be useful in the widest number of applications, a tracking algorithm needs to run close to real-time. Though our approach needs about five seconds for processing one frame, which contained six persons, but this interval is completely used for whole processes of the system, such as reading image, background subtraction step, proposing hypotheses, releasing the tracking result and learning for color appearance. Therefore, in comparing to POM+LP which requires computation for optimization, our method could run faster. In addition, the complexity of POM+LP is obviously linear with respect to grid size. Our approach does not tackle in this problem because it is only depended on the number of persons. Then, this is also the advantage of our method.

Chapter 5

Conclusion

5.1 Summarization of the thesis

In this thesis, we have presented a vision system capable of detecting multiple people, tracking them and automatically learning their color appearance. The approach was designed to work with multiple static off-the-shelf cameras.

First of all, our approach shows its advantages by using modified Bayesian model, which helps the application run faster due to the decrease in searching state space. Secondly, fusing the information of the color appearance from all camera's views by learning them in each view is another creativity. It does not only help to reduce the computation in comparing to those using a 3D shape reconstruction, but also improve the accuracy. In chapter 3 and 4, we have shown how effectively our approach used the color appearance and how it improved the tracking precision. Finally, our approach also benefited from some independent assumption such as the dynamic model, the mutual exclusion model as well as the ghost detection. This knowledge about human motivated tendency helps to smooth the trajectory of the people.

The tracking system has solved a lot of traps where other approaches had tackled. Firstly, our system was able to reliably detect and track pedestrians under reasonable crowd densities and from different viewpoints. The system has to keep in track multiple targets even in the urban environments where clutter and occlusion occur frequently. Secondly, the modified Bayesian model was very comprehensive to flexibly integrate other specific models and useful clues. We could integrate other people detection methods, such as the HOG or the optimization tracking method, into our model without any problems. Moreover, the system could be used in many applications instead of only tracking pedestrians, such as detecting the event of people falling down on the floor by extending the state space of targets. Finally, the tendency of moving such as spatial location and velocity was used to smooth the tracking trajectories.

5.2 Future work

Several future research directions have spawned from the work presented in this thesis. Among the main one is the modification of our algorithm to make it more robust to many kinds of human motion. In future work, we plan to incorporate more sophisticated appearance, such as detecting the people falling down, or jumping up. The use of an image-based person detector might help to detect each part of the human body. Moreover, it also helps in case of very crowded scenes, where people are so close to each other. In such cases, people are severely occluded on all camera views. Therefore, a part-based pedestrian detector that does not look for a whole human body but searches for isolate body parts would be recommended.

Another potential extension of our work deals with the modification of our tracking methods is to handle visibility more explicitly. Although our method tracks remarkably well even through occlusion, we believe that explicit occlusion reasoning will help to handle more difficult target interactions with missing detections, crowded scenes and long-term occlusions. Furthermore, we hope to be able to reach real-time performance with an even faster processing with more efficient implementation. This would make the method applicable to real-time application.

Efficient sampling methods also deserve attention in future research. Although the current inference method already handles challenging scenes with dozens of people, one could push toward algorithms that can deal with more challenging scenes where hundreds or thousands of people may be packed together. For example, combining MCMC with fast deterministic local search has been shown to be efficient for optimizing complex energy functions. Furthermore, improving the mixing rate of a sampler is an area of intense research. The challenge is to design efficient samplers when the topography of the underlying distribution is not known a priori, which is always the case when we are searching in a high dimensional solution space with varying dimensions, such as detecting people in crowds.

Bibliography

[1] A. Yilmaz, O. Javed, and M. Shah, “Object Tracking: A Survey,” ACM J. Computing Surveys,2006.

[2]I.Haritaoglu,D.Harwood,andL.Davis,“W4:Real-Time Surveillance of People and Their Activities,”IEEE Trans. Pattern Analysis and Machine Intelligence,vol. 22, 2000.

[3] M. Han, W. Xu, H. Tao, and Y. Gong, “An Algorithm for Multiple Object Trajectory Tracking,”Proc. IEEE Conf. Computer Vision and Pattern Recognition,2004.

[4] M. Isard and J. MacCormick, “Bramble: A Bayesian Multiple-Blob Tracker,”Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2001.

[5] T. Zhao and R. Nevatia, “Tracking Multiple Humans in Complex Situations,”IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, 2004.

[6] T. Zhao and R. Nevatia, “Tracking Multiple Humans in Crowded Environment,” Proc. IEEE Conf. Computer Vision and Pattern Recognition,2004.

[7] P. Kornprobst and G. Medioni, “Tracking Segmented Objects Using Tensor Voting,”Proc. IEEE Conf. Computer Vision and Pattern Recognition,2000.

[8] J. Berclaz, F. Fleuret, and P. Fua, “Robust People Tracking with Global Trajectory Optimization,”Proc. IEEE Conf. Computer Vision and Pattern Recognition,2006.

[9] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe, “A Boosted Particle Filter: Multitarget Detection and Tracking,”Proc. Eighth European Conf. Computer Vision,2004.

[10] G. Brostow and R. Cipolla, “Unsupervised Bayesian Detection of Independent

Motion in Crowds,"Proc. IEEE Conf. Computer Vision and Pattern Recognition,2006.

[11] Z. Khan, T.R. Balch, and F. Dellaert, "An MCMC-Based Particle Filter for Tracking Multiple Interacting Targets," Proc. Eighth European Conf. Computer Vision,2004.

[12] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking Groups of People,"Computer Vision and Image Under-standing,2000

[13] R. Rosales and S. Sclaroff, "3D Trajectory Recovery for Tracking Multiple Objects and Trajectory Guided Recognition of Actions," Proc. IEEE Conf. Computer Vision and Pattern Recognition,1999.

[14] DUANE C. BROWN. Close-Range Camera Calibration. Photogram-metric Engineering, 37(8):855–866, 1971. 20

[15] I. Sethi and R. Jain, "Finding Trajectories of Feature Points in a Monocular Image Sequence,"IEEE Trans. Pattern Analysis and Machine Intelligence,vol. 9, 1987.

[16] A. Yilmaz, X. Li, and M. Shah, "Contour-Based Object Tracking with Occlusion Handling in Video Acquired Using Mobile Cameras,"IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, 2004.

[17] J. MacCormick and A. Blake, "A Probabilistic Exclusion Principle for Tracking Multiple Objects,"Int'l J. Computer Vision,2000.

[18] Y. Wu, T.Yu, andG.Hua,"Tracking Appearanceswith Occlusions,"Proc. IEEE Conf. Computer Vision and Pattern Recogni-tion,2003.

[19] Y. Huang and I. Essa, "Tracking Multiple Objects Through Occlusions,"Proc. IEEE Conf. Computer Vision and Pattern Recogni-tion,2005

[20] A. Senior, A. Hampapur, Y. Tian, L. Brown, S. Pankanti, and R. Bolle, "Appearance Models for Occlusion Handling,"Proc. Second IEEE Workshop

Performance Evaluation of Tracking and Surveillance, 2001.

[21] N. Jojic and B. Frey, “Learning Flexible Sprites in Video Layers,” Proc. IEEE Conf. Computer Vision and Pattern Recognition,2001.

[22] H. Tao, H.S. Sawhney, and R. Kumar, “Object Tracking with Bayesian Estimation of Dynamic Layer Representations,”IEEE Trans. Pattern Analysis and Machine Intelligence,vol. 24, 2002.

[23] A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu, “Multi-Object Tracking Through Simultaneous Long Occlusions and Split and Merge Conditions,” Proc. IEEE Conf. Computer Vision and Pattern Recognition,2006.

[24] P. Kelly, A. Katkere, D. Kuramura, S. Moezzi, S. Chatterjee, and R. Jain, “An Architecture for Multiple Perspective Interactive Video,” Proc. Third ACM Int’l Conf. Multimedia,1995.

[25] K. Sato, T. Maeda, H. Kato, and S. Inokuchi, “Cad-Based Object Tracking with Distributed Monocular Camera for Security Monitoring,”Proc. Second IEEE CAD-Based Vision Workshop,1994.

[26] R. Jain and K. Wakimoto, “Multiple Perspective Interactive Video,”Proc. IEEE Int’l Conf. Multimedia Computing and Systems, 1995.

[27] A. Nakazawa, H. Kato, and S. Inokuchi, “Human Tracking Using Distributed Vision Systems,”Proc. 14th Int’l Conf. Pattern Recognition,1998.

[28] J. Orwell, S. Massey, P. Remagnino, D. Greenhill, and G. Jones, “A Multi-Agent Framework for Visual Surveillance,”Proc. IEEE Int’l Conf. Image Processing,1999

[29] Q. Cai and J. Aggarwal, “Automatic Tracking of Human Motion in Indoor Scenes Across Multiple Synchronized Video Streams,” Proc. Sixth IEEE Int’l Conf. Computer Vision,1998.

[30] J. Kang, I. Cohen, and G. Medioni, “Continuous Tracking within and across Camera Streams,”Proc. IEEE Conf. Computer Vision and Pattern Recognition,2003.

[31] T. Chang and S. Gong, “Tracking Multiple People with a Multi-Camera System,”Proc. IEEE Workshop Multi-Object Tracking,2001

[32] S. Dockstader and A. Tekalp, “Multiple Camera Fusion for Multi-Object Tracking,”Proc. IEEE Workshop Multi-Object Tracking,2001.

[33] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, “Multi-Camera Multi-Person Tracking for Easy Living,” Proc. Third IEEE Int’l Workshop Visual Surveillance,2000.

[34] A. Mittal and S. Larry, “M2tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene,”Int’l J. Computer Vision,2002.

[35] T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb, “Plan-View Trajectory Estimation with Dense Stereo Background Models,”Proc. Eighth IEEE Int’l Conf. Computer Vision,2001.

[36] A. Azarbayejani and A. Pentland, “Real-Time Self-Calibrating Stereo Person Tracking Using 3D Shape Estimation from Blob Features,”Proc. 13th Int’l Conf. Pattern Recognition,1996.

[37] S.M. Khan and M. Shah, “A Multi-View Approach to Tracking People in Crowded Scenes Using a Planar Homography Constraint,”Proc. Ninth European Conf. Computer Vision,2006

[38] J. Franco and E. Boyer, “Fusion of Multi-View Silhouette Cues Using a Space Occupancy Grid,”Proc. 10th IEEE Int’l Conf. Computer Vision,2005.

[39] D.B. Yang, H.H. Gonzalez-Banos, and L.J. Guibas, “Counting People in Crowds with a Real-Time Network of Simple Image Sensors,”Proc. Ninth IEEE Int’l Conf. Computer Vision,2003.

[40] A. Elfes, “Occupancy Grids: A Probabilistic Framework for Robot Perception and Navigation,” PhD dissertation, 1989.

[41] S. Thrun, “Learning Occupancy Grid Maps with Forward Sensor Models,”J. Autonomous Robots,2003.

[42] J. Black, T. Ellis, and P. Rosin. Multiview image surveillance and tracking. In Motion&Video Computing Workshop, 2002.

[43] J. Giebel, D. Gavrila, and C. Schn ¨ orr. ABayesian framework for multi-cue 3d object tracking. In ECCV, 2004.

[44] ROGER 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. 20, 21, 23

[45] J. Vermaak, A. Doucet, and P. Perez. Maintaining multi-modality through mixture tracking. In ICCV, 2003.

[46] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. In CVPR’08.

[47] H. Jiang, S. Fels, and J. J. Little. A linear programming approach for multiple object tracking. In CVPR, 2007.

[48] B. Leibe, K. Schindler, and L. Van Gool. Coupled detection and trajectory estimation for multi-object tracking. In ICCV, 2007.

[49] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In CVPR, 2008.

[50] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005.

[51] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. IJCV, 2005.

[52] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In CVPR, 1999.

[53] A. Andriyenko and K. Schindler. Globally optimal multi-target tracking on a hexagonal lattice. In ECCV, 2010.

[54] ZHENGYOU ZHANG. A Flexible New Technique for Camera Calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11):1330–1334, November 2000. 20

[55] J. Berclaz, F. Fleuret, and P. Fua. Multiple object tracking using flow linear programming. In Winter-PETS, 2009.

[56] STEPHEN J. MAYBANK AND OLIVIER D. FAUGERAS. A Theory of Self-Calibration of a Moving Camera. International Journal of Computer Vision, 8(2):123–151, 1992. 20

[57] MARC POLLEFEYS, R EINHARD KOCH, AND LUC VAN GOOL. Self-Calibration and Metric Reconstruction Inspite of Varying and Unknown Intrinsic Camera Parameters. International Journal of Computer Vision, 32(1):7–25, 1999. 20

[58] BILL TRIGGS. Autocalibration and the Absolute Quadric. In Con-ference on Computer Vision and Pattern Recognition, pages 609–614, 1997. 20

[59] RICHARD HARTLEY AND ANDREW ZISSERMAN. Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge, UK, second edition, 2003. 20, 21

[60] C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. IEEE Trans. on Pattern Recognition and Machine Intelligence, 22(8):747–757, 2000.

[61] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In International Conference on Pattern Recognition, volume 2, pages 28–31, 2004.

[62] Z. Yin and R. T. Collins. Moving object localization in thermal imagery by

forward-backward MHI. In IEEE Workshop on Object Tracking and Classification in and Beyond the Visible Spectrum, 2006.

[63] P. Green. Reversible jump Markov chain Monte-Carlo computation and Bayesian model determination. Biometrika, 82(4):711–732, 1995.

[64] C. Geyer and J. Mller. Simulation and likelihood inference for spatial point processes. Scandinavian Journal of Statistics, 21:359–373, 1994.

[65] Z. Tu and S. Zhu. Image segmentation by data-driven Markov Chain Monte Carlo. IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(5):657–1518, May 2002.