STEREO-BASED HUMAN DETECTION AND TRACKING FOR CROWD MONITORING

HUANG XIAOYU

(B.Eng., XMU, P.R.China)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF SINGAPORE

2004

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank my supervisor, Dr. Li Liyuan for leading me into the area of media analysis, and into the scientific research world. His concentrated and precise attitude towards research work influences me always. Without his constant encouragement and help during my stay in Institute for Infocomm Research, I could have never gone so far.

Great thanks for my co-supervisor, Assistant Professor Terence Sim from School of Computing, National University of Singapore. He has a very incisive understanding of research and he is a very good teacher. I really appreciate his inspiring guidance and invaluable advice.

I am very grateful to NUS and I^2R for offering me this opportunity to study and do research here, and for the financial support from them. The facilities and the academic staff in NUS and I^2R are the best I have ever seen.

I would like to acknowledge all my friends in Singapore. I never feel lonely having them by my side. Finally, special thanks to my family and my boyfriend for their generous love, encouragement and support for me.

TABLE OF CONTENTS

SUMMARY

In this paper, novel stereo-based methods for detecting and tracking human objects in crowds are proposed. The method for detecting human heads from a disparity image contains three distinctive steps. In the first step, Object-Orient Scale-Adaptive Filtering is proposed to extract the evidence of human heads with the most suitable scales. In the second step, a 3D virtual plane parallel and over the ground surface with the average height of human beings is built to filter out spurious evidence of human heads. Finally, a mean-shift algorithm is applied to locate human heads on the evidence map. The detected human heads are tracked by kernel-based feature evaluation, which adaptively fuses motion, color and stereo information. Tracking can provide the speed and the trajectory of each human individual in crowds, which can be used to describe the pattern of group behavior. Besides, tracking can help to reduce the error rate and increase the accuracy of the system. Good results have been achieved on the test sequences from real scenes.

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Motivation

There is a growing demand for crowd surveillance in more and more public areas such as airports, shopping malls, railway stations, museums, and stadiums. In these public sites, surveillance cameras and commercial video surveillance systems, such as Closed Circuit Television (CCTV) systems, are commonly used by security or local authorities to monitor events that involve crowd behaviors. The main aim of CCTV systems is the early detection of crowd-related unusual situations that may lead to undesirable emergencies and disasters.

Currently, the CCTV systems are mostly used to record the scenes of the latest 24 to 48 hours on the tapes for retrieving the video data “after the event”. Under this kind of circumstances, real-time surveillance cannot be achieved. For real-time monitoring with CCTV systems, crowd is monitored mostly by human operators. Undeniably, human labor is accurate within a short period, and difficult to be replaced by an automatic system. However, the limited attention span of human observers has led to significant problems in manual monitoring. The human operators are required to watch a wall of screens continuously for suspicious events, which is obviously tiring and tedious. Human labor is also costly, fatigue, and its performance deteriorates when the amounts of data to be analyzed are large. The human presence is a major limiting factor for real-time surveillance.

With the growing concern about public security, automated crowd-monitoring technique becomes very necessary. Nevertheless, the increasing reliability of high-speed processors and large memories, as well as the reducing costs of them also boost the automated crowd surveillance applications. Meanwhile, the improvement on CCTV plays an important role in accelerating the growth of automated crowd monitoring.

Group behavior understanding is crucial for automated crowd monitoring and control. Human group is made up of individuals. Hence, in order to obtain an accurate understanding of group behavior, the information about the individuals is very important. However, it remains a very challenging problem in computer vision because it is very hard to segment human individuals from the images of crowds. As a result of this difficulty, current researches in the group behavior understanding are based on the global features of the image and little has been done based on the individual information. The relationship between the global image features and the group behavior are heuristic based on the statistical results. No wonder, such methods are applicable to some simple circumstances only and give coarse estimations about crowds.

Obviously, if we can detect and track each human individual in the crowd scene, we can get better understanding of crowd behaviors and recognize unusual events for public security. It is observed that, with a fixed camera looking at an open public site from an elevated vantage point, human heads are isolated from each other in 3D space even in crowds since every human object occupies a 3D volume on the ground surface. In other words, the 3D (depth and special) evidence of human heads can be properly extracted from stereo and color image sequence, which makes it feasible to detect and track human individuals even in the very crowded scenes. Based on this information of individuals in the crowd, it is possible to develop a new methodology for group behavior understanding.

Stereo images are chosen in our research because they can provide 3D information containing both depth and spatial information of the objects in the scene. Besides, stereo images have some other distinct advantages. Stereo measure is robust under different illumination conditions and camera movements. Moreover, both stationary and moving human objects can be detected from the stereo image.

In our system, each detected person is further tracked to generate his/her spatial-temporal information. Our tracking method fuses various features, such as color, stereo and motion, for more robust tracking because each of the features may not be good enough through the sequence. Also, with tracking, the temporal information from consecutive frames can rectify the errors of head detection, which only depends on a single frame. The fusion method for tracking is a novel method since not much research has been conducted to individual detection and tracking in crowded scenes using stereo feature.

1.2 Objectives

The overall goal of this research project is to propose novel and efficient methods for human head detection and tracking, which can be used for human counting, crowd monitoring, group behavior understanding and so on.

In detail, the goal can be described as:

(1) To investigate and develop a novel method for detecting human heads in crowds from the stereo images.

(2) To investigate and develop a novel method for tracking by fusing motion, color, and stereo features.

The group behavior can be analyzed based on detecting and tracking the human individuals in the crowds. With a fixed camera looking at the crowd scene from a vantage point (e.g. camera mounted on the ceiling of a hall), the heads of human individuals are seldom occluded since each human being occupies a certain space on the ground surface. Hence the human heads are the significant isolated objects in the image sequence. Under this situation, human head detection and tracking become the fundamental functions for group behavior understanding based on the individuals of the crowd.

The disparity image can provide 3D information of the isolated human heads in the crowds. However, such data are merged with the “disparity clouds” of the human bodies and background objects. Besides, since the disparity images are not the images of dense measurement, there are many apertures from the region with little texture features in the input gray-scale images. These might give rise to false detection of heads. How to deal with such difficulties will be the main parts of our work to design a novel method to detect human heads from crowds.

When there are many human objects moving in the scene, sometimes either the visual (color) or the stereo information might be poor for the human heads observation. Hence, tracking the human heads though the sequence using just one kind of information would often lead to the loss of the targets. With the extracted stereo evidence and the visual color features of the heads, it is possible to achieve robust head tracking by adaptively fusing such two types of information.

The objectives of this research can be summarized as: to develop novel stereo-based vision methods for automated crowd monitoring, which include a method for human head detection and a method for human tracking.

1.3 Applications

The proposed human head detection and tracking method can be integrated into crowd surveillance systems. The temporal-spatial information obtained from our system, such as the motion direction, speed, and trajectory, can be used for crowd behavior analysis. For example, individuals can be clustered into groups by similar directions and velocities, and the behavior of the crowd can be analyzed further. Automated crowd monitoring systems can be developed based on our methods, and this kind of system can be used in public areas such as airport, shopping mall, railway station and so on, for the early detection of crowd-related unusual situations.

1.4 Summary of the Contributions Made

A novel stereo-based method for detecting human heads from crowds is proposed. It contains three distinctive parts: (1) object-oriented scale-adaptive filtering for feature extraction; (2) 3D perspective construction for suppressing spurious clues; (3) mean shift for head location. Most of the current crowd analysis methods only give rough global information of the target crowd. However, our method can get information of each individual in the crowd, which leads to more accurate results since crowd is made of individuals with diverse movements. The computational cost of our method is low compared to the Monte Carlo methods [28] [29]. Besides, our method is robust to variation in lighting conditions and local image pattern fluctuations to which background subtraction methods are sensitive. Most of the existing stereo-based methods are applicable to isolated person close to the camera while our method can work well with the camera mounted on a vantage point in crowded scenes where many people in groups are overlapping. Hence our stereo-based method for human head detection is a brand-new technique and it solves the problem of human individual detection in crowded scenes very well, which is proved by the experiment results.

A novel method fusing motion, color, and stereo information for head tracking is proposed. Kernels are exploited to evaluate the similarity and evidence measures on spatial, color, and stereo information. An adaptive formula is proposed to fuse the similarity and evidence measures for robust tracking in varying environment conditions. To find the accurate position of the target in the current frame is very important for tracking methods. More and more researchers are investigating in fusion methods for data association, and fusion methods are proved to provide more robust results. However, not much research has been conducted using the stereo feature in sequence with many people in groups. Our method integrates stereo information with spatial and color information for better update in the motion model, which is original and proved to be more reliable and accurate compared to those use only one or two kinds of information.

1.5 Outline of the Dissertation

The remainder of the dissertation is organized as follows:

Chapter 2 reviews the related work on crowd monitoring, stereo-based human detection, and tracking methods.

Chapter 3 presents the framework of the proposed system and the hardware configuration.

Chapter 4 describes the proposed method for human head detection from stereo images. First, the object-oriented scale-adaptive filtering is presented. Then the false clues suppressing is introduced. After that comes the head location part. The experimental results of human head detection and a quantitative evaluation are given at the end of this chapter.

Chapter 5 describes the proposed tracking method. It fuses stereo, color and motion information for robust tracking of human heads. Experimental results and evaluation for this human head tracking method are presented.

Finally, the conclusions are given in Chapter 6. In this chapter, a summary of this dissertation is made and the future work is discussed.

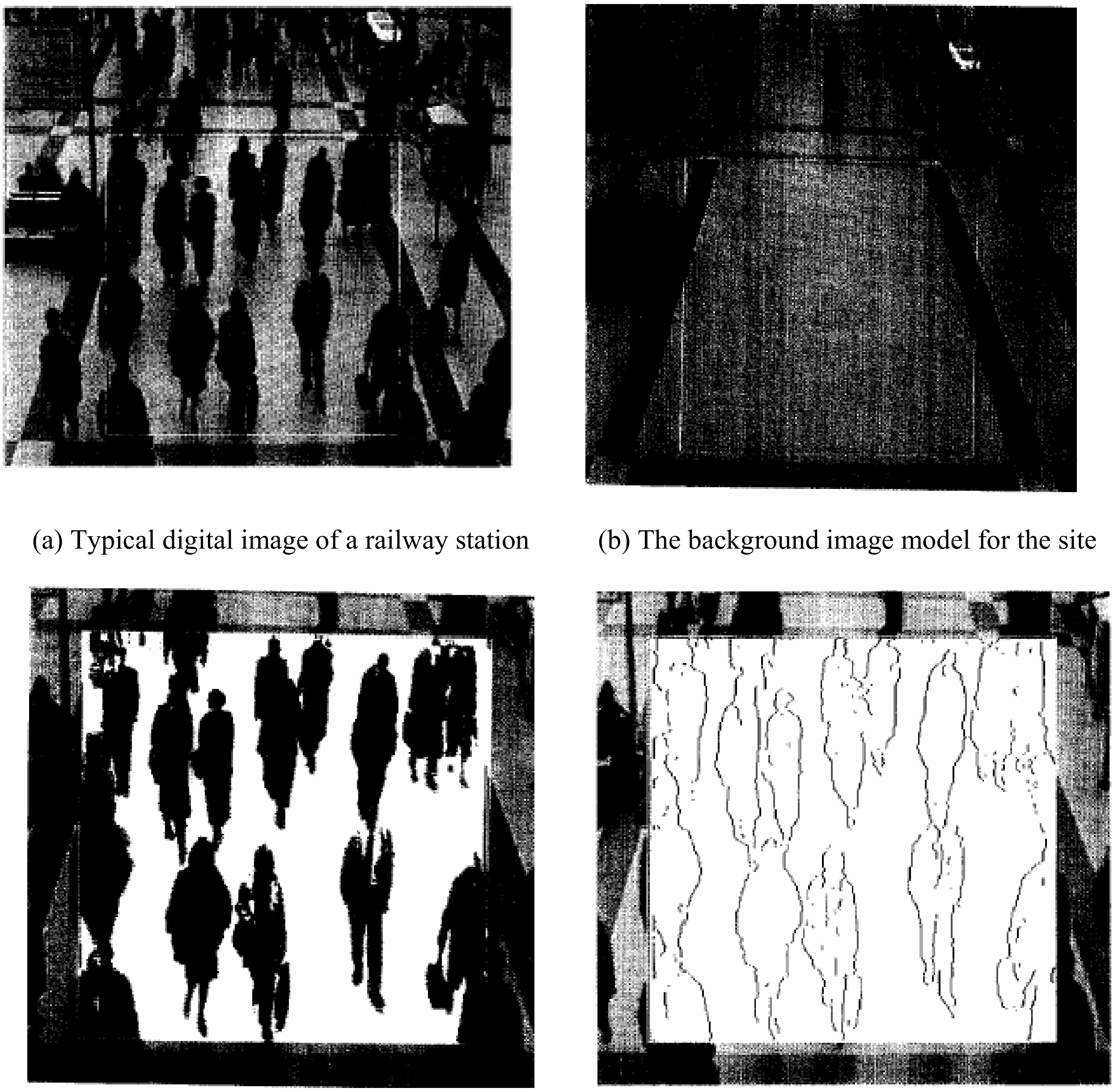
Chapter 2

Related Work

In this research, we are interested in monitoring the crowds with a stereo-based vision method. Hence, three areas of existing work are reviewed here. They are vision-based crowd monitoring, stereo-based human detection, and object tracking.

2.1 Vision-Based Crowd Monitoring

Few work has been done for vision based crowd monitoring due to the extremely difficulty of segmenting human objects from crowd scene in 2D gray-scale or color image. Instead of trying to segment each human object in the image, Velastin, *et al*, in [18], [19], [20], [6], and [17] proposed to use statistics of global image features, such as the texture, edges, and optical flows, to estimate the density of human objects in crowd scenes. One example from [6] to illustrate how to evaluate the relationship between image elements and number of people is shown in Figure 2.1, where Figure 2.1 (a) is a typical digital image of a railway station. By comparing “the crowd area” with “background area”, which can be the pixels in the image representing the crowd or the background, one can estimate the crowd density automatically.



(a) Typical digital image of a railway station

(b) The background image model for the site

(c) An image in which the background has been subtracted

(d) An image in which the edges are detected and thinned

Figure 2.1 An example of the effect of the background subtraction and edge detection

Here background removal is applied first, and then quantitative measures are performed. Figure 2.1 (b) is a "background-only" image for the site, and Figure 2.1 (c) is the result of applying background removal to Figure 2.1 (a).

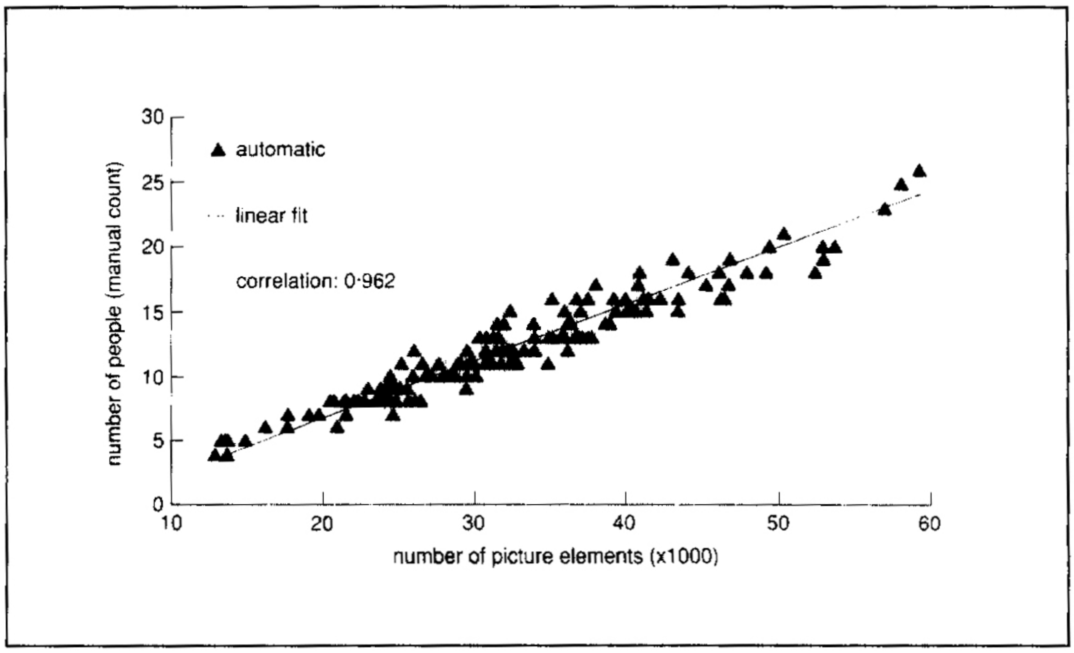


Figure 2.2 The relationship between the number of people and the pixel number of crowd obtained by background subtraction in an image

Without counting the pixels of the crowd, one can measure the perimeter of the regions of the crowd as an alternative. However, the thickness of the edges varies which affects the accuracy. Hence, the edges are thinned in advance for better results. A typical image containing thinned edges is shown in Figure 2.1 (d), and the relationship between the number of people and the number of edge pixels is shown in Figure 2.2. The edge detection method is a standard low-level function for image processing, and it is efficiently in some situations, but occlusion and overlapping of individuals will bring much error in the edge detection method. The relationship between the number of people and the pixel number of extracted and thinned edges for human beings in an image is shown in Figure 2.3.

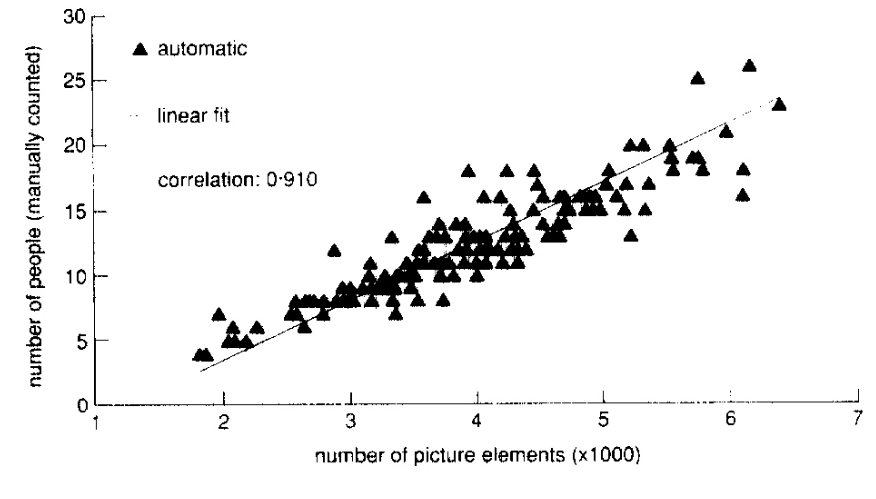


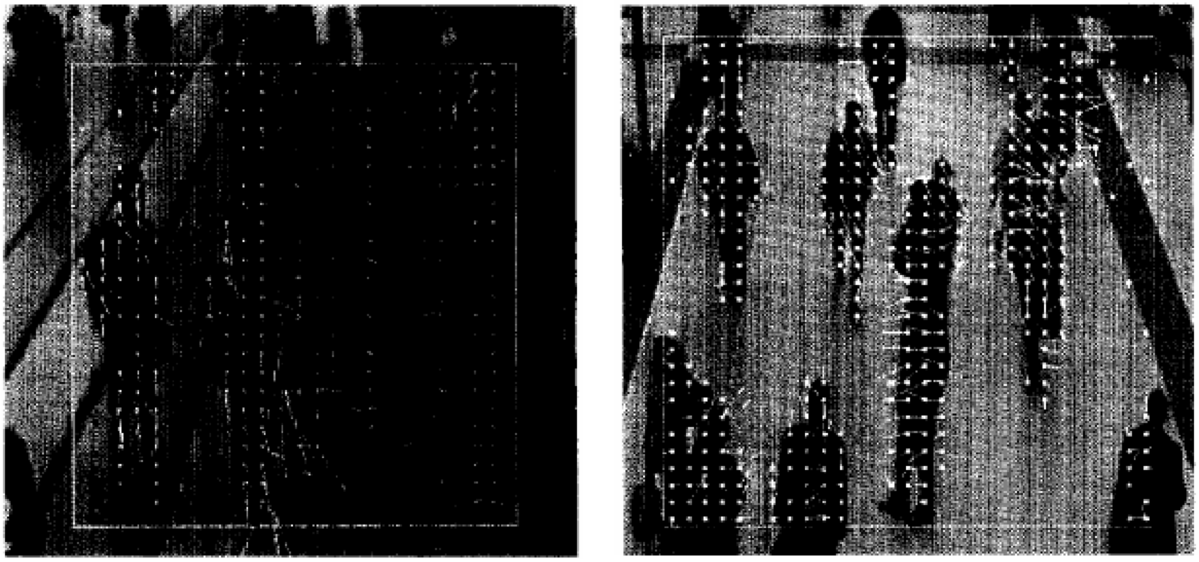
Figure 2.3 The relationship between the number of people and the pixel number of extracted and thinned edges for human beings in an image

With the relationship between the image elements of crowd and the number of people, one can estimate the density of the crowd in the image. Trained classifiers, e.g., neural network, are used to classify the scene into two to five categories, such as low, high, or very high of person densities [20]. However, only the rough estimation of crowd density can be obtained by using this method.

In [6], A. Devies, J. Yin and A. Velastin tried to use some image processing techniques to estimate crowd motion. Aware of the difficulty of detecting and tracking individuals in image sequences, they chose to find general models for crowd analysis, which do not rely on detecting individuals. The optical flow method is investigated in this work. The optical flow is a vector corresponding to the change of image brightness between consecutive frames. Let $I(x,y,t)$ denote the image brightness of (x,y) at time *t*, then the optical flow is calculated as:

$$\frac{dI}{dt}=\frac{\partial I}{\partial x}\frac{dx}{dt}+\frac{\partial I}{\partial y}\frac{dy}{dt}+\frac{\partial I}{dt}=\frac{\partial I}{\partial x}u+\frac{\partial I}{\partial y}v+\frac{\partial I}{dt} \tag{2. 1}$$

What we need is the motion vectors (u,v). The “optical flow constraint” (OFC) is a well-known method to solve the above equation, which assumes that the brightness is constant with regard to time (i.e. $dI/dt=0$). However, this assumption is difficult to satisfy when occlusion or sudden changes in illumination happen. The typical result of crowd movement detection is shown in Figure 2.4 (a).



(a) Movement detected by "optical flow" (b) Movement calculated by "block matching"

Figure 2.4 The typical results of motion calculated by "optical flow" and "block matching"

They also investigated the Block-matching motion detection method. An object block is selected in the current frame, and then to find the object in the second frame is to find the same-sized block with the highest similarity within the "search area" in the next frame. After that, the motion vector of the crowd, including the velocity and direction is defined by the displacement of the object block. Figure 2.4 (b) demonstrates the result of a motion calculation by "block matching". These crowd analysis methods only give global information of the monitored crowd. In addition, when there is a great diversity of individual movements in the crowd, the density estimated by this kind of methods are not accurate enough.

In [30], Monique Thonnat, *et al*, propose an approach to recognize human behavior for metro surveillance. In their system, 3D scene models for all the cameras are constructed respectively. Then they apply the Motion Detector to detect the moving regions in every frame and further classify them into various types, such as *PERSON, GROUP, TRAIN* and so on. After a list of moving objects is obtained, a Frame to frame tracker is used to link the moving objects through the consecutive frames. They also combine all the graphs of the moving objects from all the calibrated cameras viewing the same scene with overlapping fields into a global graph for better results. This system can track human targets in three ways: Individual, Group of people and Crowd. One of the ways is chosen by the system depending on the situation of the scene for more accurate and efficient processing. Finally, behavior recognition is performed and the description of the current scene is made for metro surveillance. The human densities of the crowds are estimated by using the method proposed by Velastin as reviewed previously.

Recently, Zhao and Nevatia proposed a method to segment human individuals in foreground regions obtained by background subtraction [28]. They try to interpret the foreground region with a configuration of a number of human individuals by maximizing the posterior probability. Specific knowledge about human shape, height, camera setting and image cues of head contours are integrated in a Bayesian framework. They also achieved multiple humans tracking [29] based on the above research. In this system, they employed the Markov chain Monte Carlo (MCMC) to compute the optimal solution, and they used various direct image features to make the Markov chain work efficiently. In their work, camera models as well as human shape models are required to be known in advance. Thousands of iterations are required to get an optimal solution for a frame containing many persons.

2.2 Stereo-Based Human Detection

Stereo images can provide 3D information of the objects in the scene. Hence, it should be easier to detect the isolated human objects in 3D space from the stereo images than to segment them from 2D color images, especially when there are partial occlusions of persons from the viewpoint of the camera. However, since generating the stereo image by matching the left and right images from a calibrated stereo head requires dense computation, only after the late of 90's there have been reports of employing stereo images for human detection and tracking from a video.

I. Haritaoglu, D. Harwood and L. Davis proposed a real-time surveillance system for finding and tracking people in the scenes, i.e., W^4S in [9]. It integrated the stereovision into W^4 [10], which is their earlier work only depending on 2D gray scale images. In W^4S, firstly both intensity-based and stereo-based detection algorithms are applied for the foreground object detection. Then areas, in which there are significant overlaps between foreground regions detected by disparity and by intensity, are confirmed as foreground regions. A foreground object is defined as the intersection of a connected component, which contains only one intensity region and one disparity region that have very high overlap. As to tracking, W^4S applies a second order motion model for each object to predict its position, followed by a two-stage matching method for the model update. The first matching stage is the initial estimation of object displacement, which is computed as the motion of the median of the object coordinates. The second stage is a binary edge correlation between the current displaced the object and that in the previous frame based on their silhouette profiles. This stage makes the estimation more accurate than the first stage. The overall performance of the silhouette-based technique relies heavily on the accuracy of foreground segmentation, and the cardboard model used inW^4S to predict human pose and position is restricted to upright people. It is designed to track a single isolated pedestrian in the scene.

C. Eveland, K. Konolige and R.C. Bolles presented a method based on a statistical model of a stereo background image to track people and detect human heads in a stereo image sequence [7]. Gated Background Adaptation, the dynamic version of the model, is able to extract background statistics from a stereo video stream in the presence of corrupting foreground objects. The discrimination of background and foreground is the first stage for human detection and tracking. After that, foreground blobs with the similar shapes to human head are extracted as potential tracking targets. Finally, a simple velocity model is applied to track an acquired target. A stereo head mounted on a pan/tilt platform follows the tracked targets to keep them in the field of view. Another real-time tracking method for multiple people tracking is proposed in [1]. In this method, background differencing is applied to the input stereo images to detect the foreground objects in the scene. Firstly, the background is learned when the scene contains no people. Then, pixels with a larger disparity than that of the background are initially defined as foreground. After noise filtering, a histogram of the foreground disparity image is constructed and smoothed. The peaks in the histogram are used as seeds to segment the foreground into layers of near constant disparity. Correlation with binary templates of the 2D shapes of human being is applied to detect people in the foreground layers. The tracker is a constant velocity Kalman filtering model where the measurement process is intensity correlation with adjustments from stereo detection.

As to the background subtraction methods as well as background differencing methods, when the scene is crowded with many overlapping persons in groups, it becomes difficult to detect human individuals from the foreground regions. In Addition, occlusion problem becomes serious in the high-density situation and leads to low accuracy [25]. The result of this method is also easily affected by the variations of overall lighting conditions and local image pattern fluctuations. Some amendment has been made to the method but hardly deals with sudden or major changes in the background. The above problems remain as the inherent turns out to be the bottleneck of the background subtraction methods. There are other existing stereo-based methods developed to track isolated human object close to the camera for human machine interaction [21], [8], [24], [7]. For example, a head tracking system in [24] using stereo depth information together with a simple human torso model is claimed to be fast and robust. Russakoff and M. Herman constructed a depth model of the background for the foreground segmentation. Then they applied edge detectors on the foreground to find occluding edges as features for a simple torso model fitting. In this case, the distance from the object to the camera is less than 5 meters and there is only one object in the scene. The limitations are similar as to the system presented in [5] by T. Darrell, *et al*. The stereo, color, and face detection modules are integrated into a single framework. If a person does not turn his face to the camera, this system cannot recognize him. The application is restricted to short distance from the camera to the object because if a person is far from the camera, the face in the image will be too small to be detected by face pattern, and the skin color area will be insignificant. Also, this system cannot detect peoples in crowd, since the camera height is about the human height and people in the back of the crowd will be seriously occluded by those in the front.

When the scene is crowded with many overlapping peoples in groups, it becomes very difficult to detect human individuals using these methods because of the serious occlusion problem and the long distance from the crowd to the camera.

2.3 Object Tracking

Visual tracking is an important part for human behavior analysis. Tracking can provide spatial-temporal information, such as motion speed, direction, and trajectory for human behavior understanding.

Under the traditional formation, tracking is thought as a probabilistic inference problem. Given a sequence of previous observations $\mathbf{y}_0,\cdots,\mathbf{y}_{i-1}$, it tries to find the current observation $\mathbf{y}_i$ with the maximum posterior probability

$$P(\mathbf{y}_i \mid \mathbf{y}_0,\cdots,\mathbf{y}_{i-1}) \tag{2. 2}$$

If a motion model is employed to describe the internal state of the target as $\mathbf{X}_i$, Equation (2. 2) can be written as

$$P(\mathbf{y}_i \mid \mathbf{y}_0,\cdots,\mathbf{y}_{i-1}) = \int P(\mathbf{y}_i \mid \mathbf{x}_i) P(\mathbf{x}_i \mid \mathbf{y}_0,\cdots,\mathbf{y}_{i-1}) d\mathbf{x}_i \tag{2. 3}$$

where $P(\mathbf{x}_i \mid \mathbf{y}_0,\cdots,\mathbf{y}_{i-1})$ is the prediction based on the motion model and $P(\mathbf{y}_i \mid \mathbf{x}_i)$ measures the observation (data association). The existing methods for prediction and data association will be reviewed separately in the following two subsections.

Prediction: In general, the existing motion models can be divided into two classes: linear and non-linear dynamic models.

Assuming that all the conditional probabilities are normal, which means $P(\mathbf{x}_i \mid \mathbf{y}_0,\cdots,\mathbf{y}_{i-1})$ is normal, it becomes relatively simple to update the representation of the mean and covariance for the prediction. Let's use the notation $x \sim N(\mu,\Sigma)$ to mean that x is the value of a random variable with a normal probability distribution with mean μ and covariance Σ. So the linear dynamic models can be written as

$$\mathbf{x}_i \sim N(\mathbf{D}_i\mathbf{x}_{i-1},\Sigma_{\mathbf{d}_i}) \tag{2. 4}$$

$$\mathbf{y}_i \sim N(\mathbf{M}_i\mathbf{x}_i,\Sigma_{\mathbf{m}_i}) \tag{2. 5}$$

Where x_i is the estimate position of the object at Frame *i*; $\mathbf{y}_i$ is the value of the measurement at Frame *i*; $\mathbf{D}$ is the dynamics matrix and $\mathbf{M}$ is the measurement matrix.

Among Linear dynamic tracking models, Kalman filtering has been a very popular one [14], [11], [2]. From Equation (2. 4) and Equation (2. 5), the dynamic model for a 1D State Kalman Filter can be written as:

$$\mathbf{x}_i \sim N(\mathbf{d}_i\mathbf{x}_{i-1}, \sigma^2_{\mathbf{d}_i})$$

$$\mathbf{y}_i \sim N(\mathbf{m}_i\mathbf{x}_i, \sigma^2_{\mathbf{m}_i}) \qquad (2.\ 6)$$

Kalman filter uses a form of feedback control to maintain tracking on an object. The approach works by iteratively propagating a Gaussian state density function described by its mean and a covariance matrix. At the start of every iteration, we assume that $P(\mathbf{x}_{i-1} | \mathbf{y}_0, \cdots, \mathbf{y}_{i-1})$ is known. Let's represent the mean and standard deviation of $P(\mathbf{x}_i | \mathbf{y}_0, \cdots, \mathbf{y}_{i-1})$ as $\overline{\mathbf{x}}_i^-$ and σ_i^- respectively, and the mean and standard deviation of $P(\mathbf{x}_i | \mathbf{y}_0, \cdots, \mathbf{y}_i)$ as $\overline{\mathbf{x}}_i^+$ and σ_i^+ respectively. When tracking a one-dimensional state variable using the dynamic model in Equation (2. 6), the 1D Kalman filter updates predictions of the mean and covariance.

The prediction equations are updated by

$$\overline{\mathbf{x}}_i^- = \mathbf{d_i}\overline{\mathbf{x}}^+_{\mathbf{i-1}} \qquad (2.\ 7)$$

$$\sigma_i^- = \sqrt{\sigma_{\mathbf{d}_i}^2 + (\mathbf{d}_i \sigma_{i-1}^+)^2} \qquad (2.\ 8)$$

and the correction equations are updated as

$$\mathbf{x}_i^+ = \left(\frac{\overline{\mathbf{x}}_i^- \sigma_{\mathbf{m}_i}^2 + \mathbf{m}_i \mathbf{y}_i (\sigma_i^-)^2}{\sigma_{\mathbf{m}_i}^2 + \mathbf{m}_i (\sigma_i^-)^2} \right) \qquad (2.\ 9)$$

$$\sigma_i^+ = \sqrt{\left(\frac{\sigma_{\mathbf{m}_i}^2 (\sigma_i^-)^2}{\sigma_{\mathbf{m}_i}^2 + \mathbf{m}_i^2 (\sigma_i^-)^2} \right)} \qquad (2.\ 10)$$

Kalman filtering is very useful in cases that tracking needs to be done under very noisy conditions. The disadvantage of the Kalman filter is that it relies on a good initial estimate for convergence. In Addition, Kalman filter will fail in tracking objects with complex dynamics. The fundamental reason for this inadequacy is the Gaussian representation of probability density, while the presence of background clutter and self-occlusions will always lead to multi-modal density of the state space.

There is always only one peak in the posterior in a linear dynamic model with linear measurements. However, a substantial number of peaks may appear even if the dynamic models have very insignificant non-linearity, and many natural dynamic applications are non-linear. Linear dynamic models may fail in this kind of cases. Hence non-linear dynamic models are developed to solve the problem. It can be written as:

$$\mathbf{x}_i \sim N(f(\mathbf{x}_{i-1}, i); \Sigma_{\mathbf{d}_i}) \tag{2. 11}$$

where f is a non-linear function. Neither $P(\mathbf{x}_i \mid \mathbf{y}_0, \cdots, \mathbf{y}_{i-1})$ nor $P(\mathbf{x}_i \mid \mathbf{y}_0, \cdots, \mathbf{y}_i)$ is normal, and $P(\mathbf{y}_i \mid \mathbf{x}_i)$ may not be Gaussian, either. As one of the non-linear dynamic models, Particle filter works well in some applications under this kind of situations [26], [23]. A standard Particle filter is introduced as follows.

What we need to do is to obtain a representation of $P(\mathbf{x}_i \mid \mathbf{y}_0, \cdots, \mathbf{y}_i)$ from a sampled representation of $P(\mathbf{x}_i \mid \mathbf{y}_0, \cdots, \mathbf{y}_{i-1})$. Assume that we have a collection of N points u_i^n sampled from an appropriate proposal distribution $q(\mathbf{x}_i \mid \mathbf{x}_0, \cdots, \mathbf{x}_{i-1}, \mathbf{y}_0, \cdots, \mathbf{y}_{i-1})$, and a collection of weight w_i^n in the simplest case. $q(\mathbf{x}_i \mid \mathbf{x}_0, \cdots, \mathbf{x}_{i-1}, \mathbf{y}_0, \cdots, \mathbf{y}_i)$ could be set as $p(\mathbf{x}_i \mid \mathbf{x}_{i-1})$. Then the importance ratios for these particles are calculated as:

$$w_i^n = w_{i-1}^n \frac{p(\mathbf{y}_i \mid u_i^n) p(u_i^n \mid \mathbf{x}_{i-1}^n)}{q(u_t^n \mid \mathbf{x}_0^n, \cdots, \mathbf{x}_{i-1}^n, \mathbf{y}_0, \cdots, \mathbf{y}_i)} \tag{2. 12}$$

Using the importance weights, we resample the particles to generate an un-weighed approximation of the posterior distribution $P(\mathbf{x}_i \mid \mathbf{y}_0,\cdots,\mathbf{y}_{i-1})$. In the mixture approach, the particles are used to obtain the $P(\mathbf{x}_i \mid \mathbf{y}_0,\cdots,\mathbf{y}_{i-1})$ as:

$$P(\mathbf{x}_i \mid \mathbf{y}_0,\cdots,\mathbf{y}_{i-1}) \approx \sum_{j=1}^{N} \Pi_{j,i} \sum_{n\in \mathrm{I}_j} w_i^n \delta_{\mathbf{x}_i^n}(\mathbf{x}_i) \qquad (2.\ 13)$$

where I_j is the indices set of the particles of the *j*-th mixture component.

As mentioned above, non-linear dynamic models are applicable to non-linear dynamics such as non-Gaussian densities, and they maintain multiple samples to provide multiple hypotheses. However, the prior density has to be predetermined (from Kalman filter or other methods). In Addition, if the number of the required samples *N* is very large, the required number of samples will grow exponentially with the size of the state space, which will lead to expensive computational cost and big latency in processing.

Data association: Determining which measurements are informative about the state of the object being tracked is always called data association. Probably, data association is the biggest source of difficulties in applications in computer vision. Using those informative measurements to guide the matching part of our tracking process always increase the accuracy in updating the representations of the status of the objects being tracked.

Methods based on various features for visual tracking have been investigated. Basically, they are dependent on various visual features, such as color distribution, texture, and contours. As to color-based methods for matching, some system just use the color difference between the original object area and the candidate object area while some use more sophisticated measures such as color distributions. In [22], K. Nummiaro, *et al* integrated adaptive color distributions into particle filtering. In the data association stage, a distance function is applied first to suppress the information of the pixels that are far away from the region center:

$$k(r) = \begin{cases} 1 - r^2 & : r < 1 \\ 0 & : otherwise \end{cases} \tag{2. 14}$$

where r is the distance from the region center. Suppose the color space has m bins. The color histogram of a region R at location y is defined as

$$p_y^{(u)} = f \sum_{x_i \in R} k\left(\frac{\| y - x_i \|}{a}\right) \delta[h(x_i) - u] \tag{2. 15}$$

where δ is the Kronecker delta function and $h(x_i)$ represents a given color at location x_i in the m-bins histogram. f is the normalization factor that assures $\sum_{u=1}^{m} p_y^{(u)} = 1$ and a makes the system invariant against the region scaling. The Bhattachayya coefficient is used to measure the similarity between the candidate object and the original object. The coefficient of two color histograms $p = \{p^{(u)}\}_{u=1...m}$ and $q = \{q^{(u)}\}_{u=1...m}$ is calculated as

$$\rho[p,q] = \sum_{u=1}^{m} \sqrt{p^{(u)} q^{(u)}} \qquad (2.\ 16)$$

The larger ρ means the larger similarity between the two distributions. The Bhattachayya distance between these two distributions is defined as

$$d = \sqrt{1 - \rho[p,q]} \qquad (2.\ 17)$$

The methods based on color distributions would fail to track a head that changes its pose during motion. The contour-based methods for matching are also popular. Generally an edge detector or a contour model is applied to the images first, and then the contour in the current frame that is closest to the object contour in the previous frame is selected as the tracked object. A good example of the contour-based tracking is proposed in [13]. In the matching process, measurement *z* for the object curve *x* represents edge fragments found by edge-detector. The noise and distortions are assumed to be local, to narrow the image pixels for examination to those near the image curve. Use $x(s), 0 \leq s \leq 1$ to denote the image curve *x*, and $z(s)$ to denote the corresponding measurement sequence *z*, which are the detected edges closest to x along curve normals. The measurement density is formulated as a truncated Gaussian as the following:

$$p(z \mid x) = \exp\left\{ -\frac{1}{2\sigma^2} \int_0^1 \phi(s) ds \right\} \tag{2. 18}$$

where

$$\phi(s) = \begin{cases} | x(s) - z(s) |^2, & if \mid x(s) - z(s) \mid < \delta \\ \mu, & otherwise \end{cases} \tag{2. 19}$$

where μ is a penalty constant for failure in finding a feature and δ is to confine the area for matching. Features of points with distances to the curve larger than δ are not used. The smaller ϕ will lead to the bigger $p(z \mid x)$, which means the similarity between the observed and the tracked object is bigger. Contour-based methods would fail in clutter images, including crowds and complex background.

To find the closest position for the object tracked in the image, various image cues have been proposed in matching, such as the color-based and contour-based methods. Because each of the features may not be robust enough when working alone, more and more research has been conducted to integrate multiple visual cues [3]. Fusion methods for tracking have been proved to provide more robust results. However, not much research has been conducted to individual detection and tracking in crowded scenes with stereo images. Our method for head tracking fuses the information of stereo, color and motion for data association. It is more reliable and accurate compared to those use only one or two kinds of information, which is proved by the experiment results.

2.4 A Brief Comparison Between the Existing Methods and Our Methods

To illustrate the difference between the existing methods reviewed above and our methods, a brief comparison is made in Table 2.1. The comparison is made with respect to features exploited and objects described. The features can be clustered into three types, i.e., visual information (color or contour), stereo measures, and motion estimation. The described objects could be the isolated individuals or crowds. It can be seen from the table that our method can get information of each individual in the crowd and our method is robust to variation in lighting conditions and local image pattern. Most of the existing stereo-based methods are applicable to isolated person close to the camera while our method works well with serious occlusion in the crowded scene. Hence our stereo-based method for human head detection is a brand-new technique and it solves the problem of human individual detection in crowded scenes very well, which is proved by the experiment results.

Existing & Our methods	Features used			Objectives	Notes
	Stereo	Visual	Motion		
Velastin [6]	×	√	√	Get the information of the crowd as a whole	Occlusion leads to much error in this system. Can only obtain rough global information of the crowd
Zhao [28] [29]	×	√	√	Detect and track individuals using Bayesian framework	High Computational Cost; Specific knowledge required
Haritaoglu [9] (W^4S)	√	√	√	Detect and track upright isolated human objects	Sensitive to the change in the background.
Konolige [7] [1]	√	√	√	Detect and track isolated person close to the camera	Sensitive to the change in the background; Knowledge of background Required.
Russakoff [24]	√	√	√	Construct a Human-Machine	Limited to isolated person close to the camera
Darrell [5]	√	√	√	Construct a Human-Computer Interface	Limited to isolated person close to the camera
Ours	√	√	√	Detect and track individuals in crowds with a distance as 10~30m to the camera	Can not distinguish between human head and non-human object with shape and height similar to those of human; Field of stereo vision is limited

Table 2.1 A Brief Comparison between Existing methods and Our methods

Chapter 3

Framework and Configuration

3.1 Framework of Proposed System

We propose to use a stereo head to monitor the crowd scene in an open area from an elevated vantage point. The general framework for the proposed system is illustrated in Figure 1. As the flow chart shows, the input is the stereo and color image and the output would be the recognized patterns of group behaviors. There are four consecutive processing modules:

- **Stereo Image Generation:** Run a set of functions in the library provided by The Small Vision System (SVS) to capture images from the left and right cameras, and generate the stereo image (disparity image).

- **Human Head Detection:** Extract the evidence of human heads with a scale-adaptive filtering, suppress the spurious clues and locate the positions for the true heads.

- **Human Head Tracking:** Track each human individual in the crowd though the sequence by fusing the information of motion, color and stereo. Generate the speed and trajectory for human motion description.

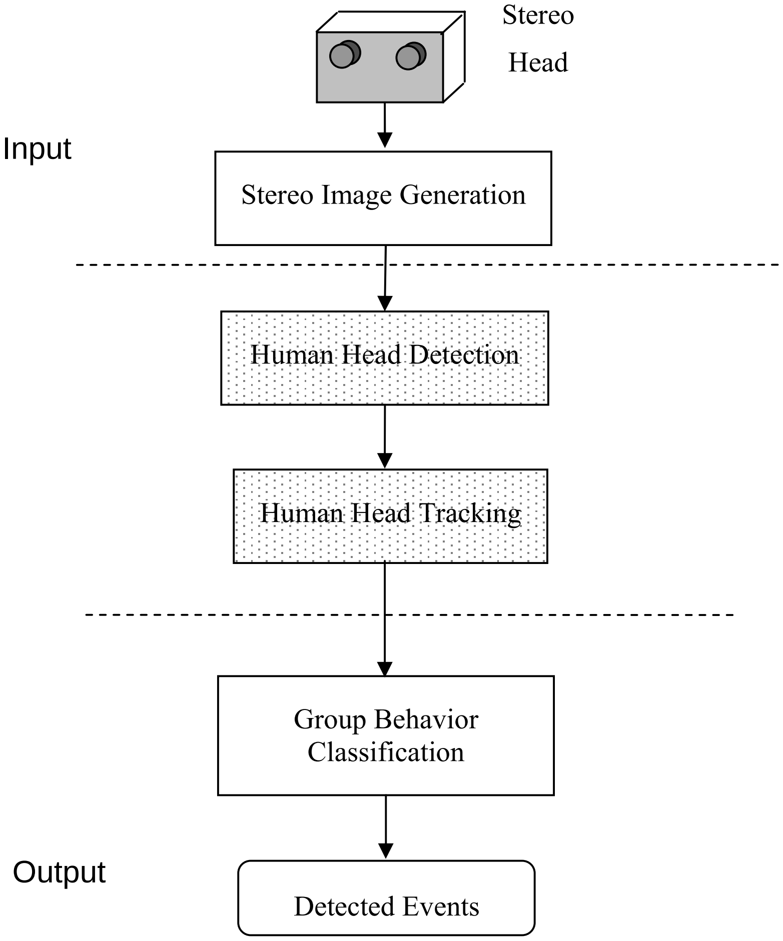


Figure 3.1 System Overview (The two shaded blocks in the middle are the main modules developed in this thesis)

- **Group Behavior Classification:** Classify the motion of crowds into a few patterns by clustering the individuals into groups based on their motion features. This will be done in the future.

3.2 Hardware Configuration

In this research, a stereo head from The Small Vision System (SVS) of SRI is used to generate stereo images. SVS is an implementation of the Stereo Engine, which can be apply on PCs [15]. The vendor provides a set of library functions implementing the stereo algorithms. In our program, some of these functions are called to capture images by the camera and compute stereo image from a pair of left and right images. The configuration of the SVS system is shown in Figure 2.

In this work, a stereo head with wide baseline is used. The type of the stereo head is Videre Design MEGA-D STH-MD1/-C. The baseline of the stereo head is 30cm. The focal lengths of the left and right cameras are 16mm. The physical appearance of the stereo head is shown in Figure 3. The distance from the camera to the scene is about 10m~30m while the height of the stereo head from the ground surface is about 10m.

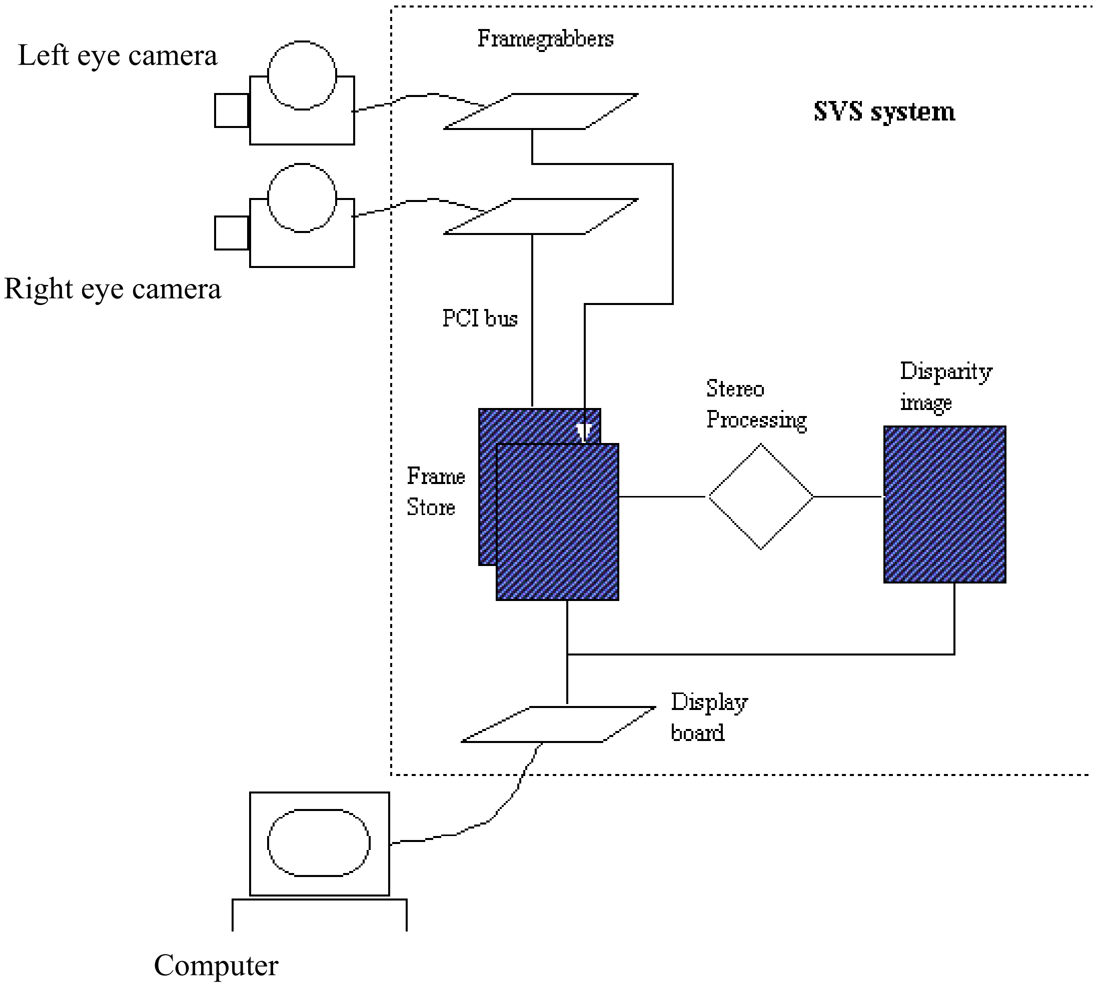


Figure 3.2 The development environment of the Small Vision System.

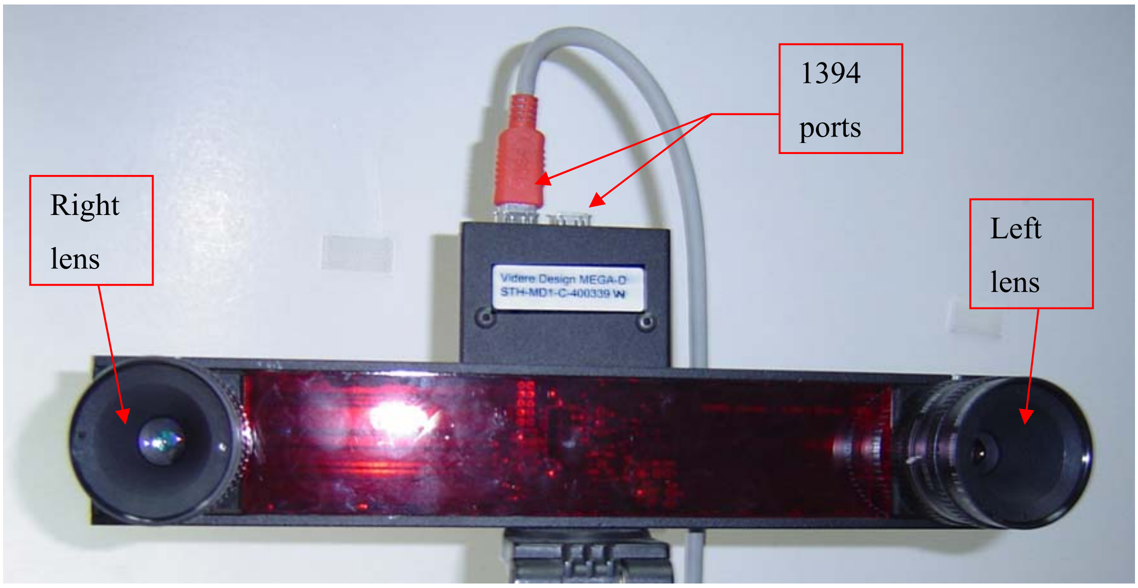


Figure 3.3 Physical appearance of the stereo head.

Chapter 4

Human Head Detection From Stereo Images

To analyze the crowd behavior based on the motion of individuals, the first task is to detect the individuals. Here, human heads are used as targets for individuals in crowds since they could be observed significantly from an elevated vantage point. The proposed method [12] to detect human heads from stereo images contains three steps: (1) scale-adaptive filtering [16], (2) false clue suppressing, and (3) head location. They are described in detail in the following sections. Experiments on real images and short summary are presented in the last two sections.

4.1 Object-Oriented Scale-Adaptive Filtering

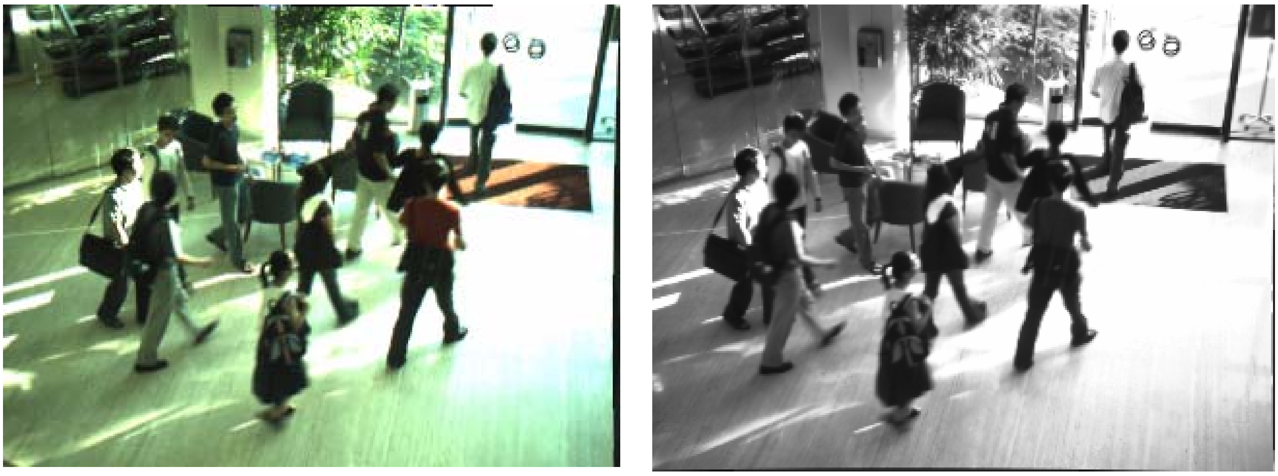
The stereo image is a gray-scale image where the intensity represents the disparity value at the pixel. The relation between the disparity (*d*) and the depth (*z*) to the camera is

$$d = \frac{bf}{z} = \frac{K_1}{z}$$

or

$$z = \frac{bf}{d} = \frac{K_1}{d} \qquad (4.\ 1)$$

where b is the baseline and f is the focal length of the lens. $K_1 = bf$ is a constant for each stereo camera. After calibration, the disparity image $g(x, y)$ can be obtained from the left and right images captured by the stereo camera. Examples of the color and disparity images are shown in Figure 4.1 and Figure 4.2 respectively.



(a) Image of the left camera

(b) Image of the right camera

Figure 4.1 One example of images from the stereo head

The stereo information allows us to estimate the size of a human head with the corresponding distance to the camera. Let $d = g(x, y)$ be the disparity value at the pixel (x, y). If it is the center of a head with the distance $z = K_1 / d$ to the camera, the corresponding disparity values from the head will be within the range of $[d_-, d_+](d_- < d < d_+)$ with

$$d_{+} = \frac{K_1}{z - \frac{D_h}{2}} = \frac{K_1 d}{K_1 - \frac{D_h d}{2}}$$

$$d_{-} = \frac{K_1}{z + \frac{D_h}{2}} = \frac{K_1 d}{K_1 + \frac{D_h d}{2}} \qquad (4.\ 2)$$

where D_h is the average depth of human heads. An example is illustrated in Figure 4.3, in which two individuals stand with different depths z_1, z_2 to the camera. The depth ranges of the heads of them are $[z_{10}, z_{11}]$ and $[z_{20}, z_{21}]$ respectively.

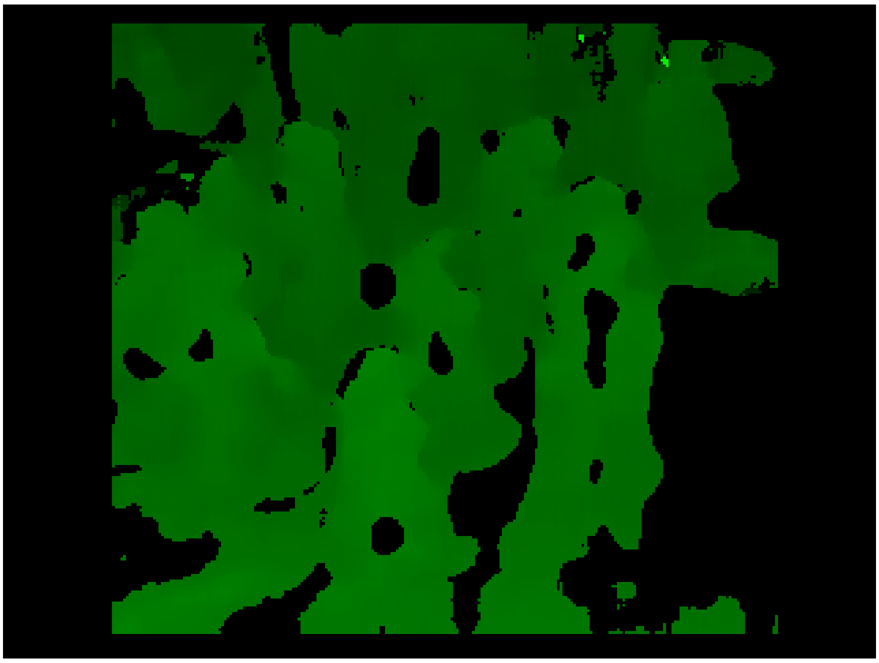


Figure 4.2 The disparity image generated from the example in Figure 4.1

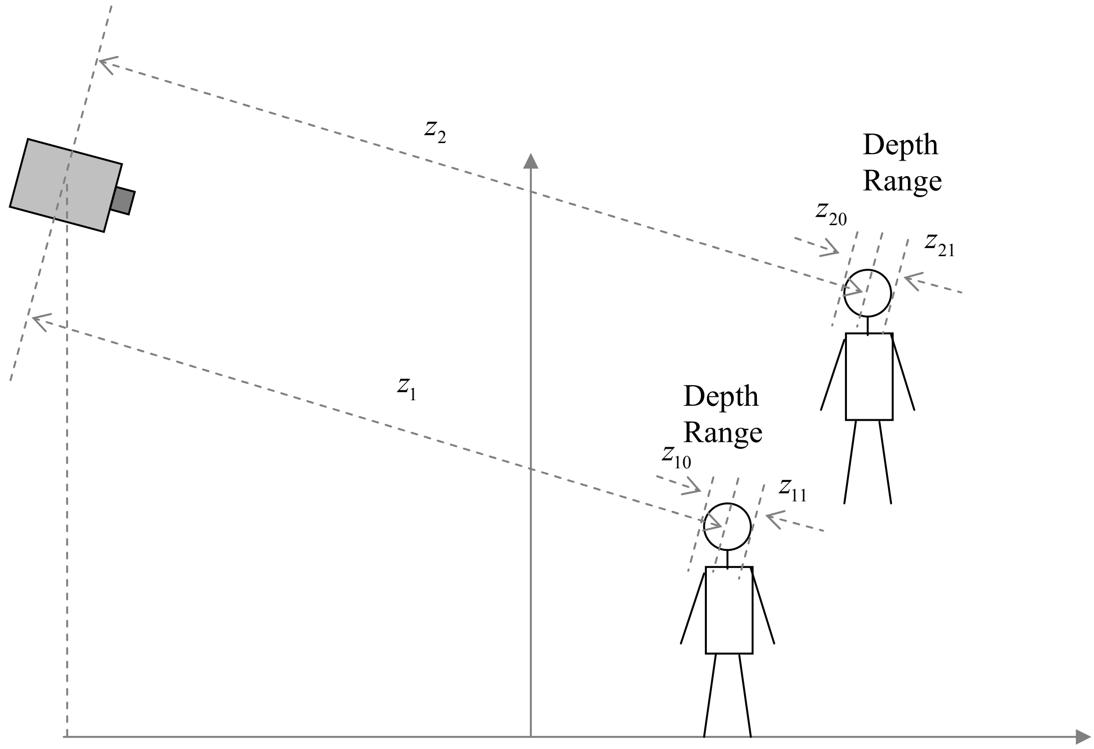


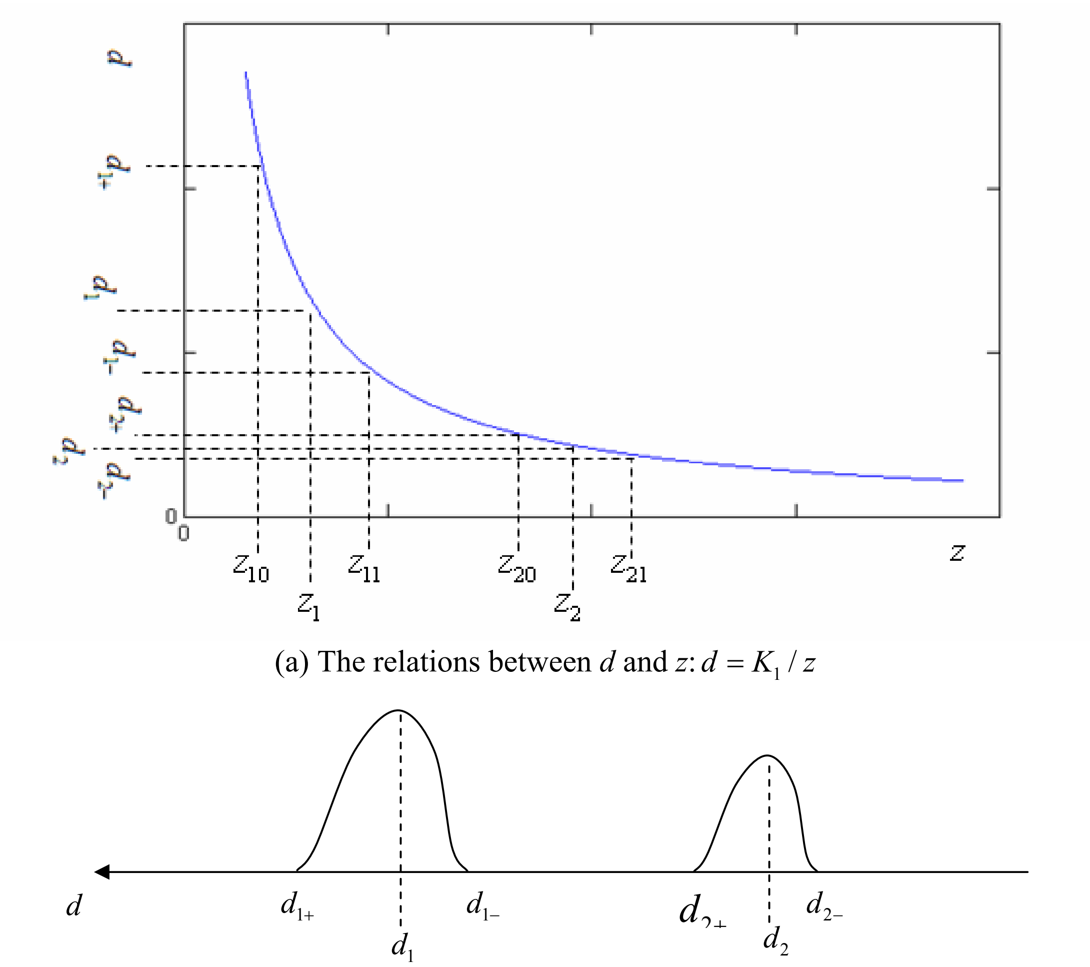
Figure 4.3 Human heads with different depths to the camera

The corresponding relationship between *z* and *d* for the example is shown in Figure 4.4(a) and the distributions of the disparity measures for the two persons are shown in Figure 4.4 (b). From the simple perspective triangles as shown in Figure 4.5 (a) and (b), the estimated width and height of the head in the image can be obtained as

$$w_h(d) = \frac{fW_h}{z} = \frac{K_2 W_h}{d}$$

$$h_h(d) = \frac{fH_h}{z} = \frac{K_2 H_h}{d} \qquad (4.\ 3)$$

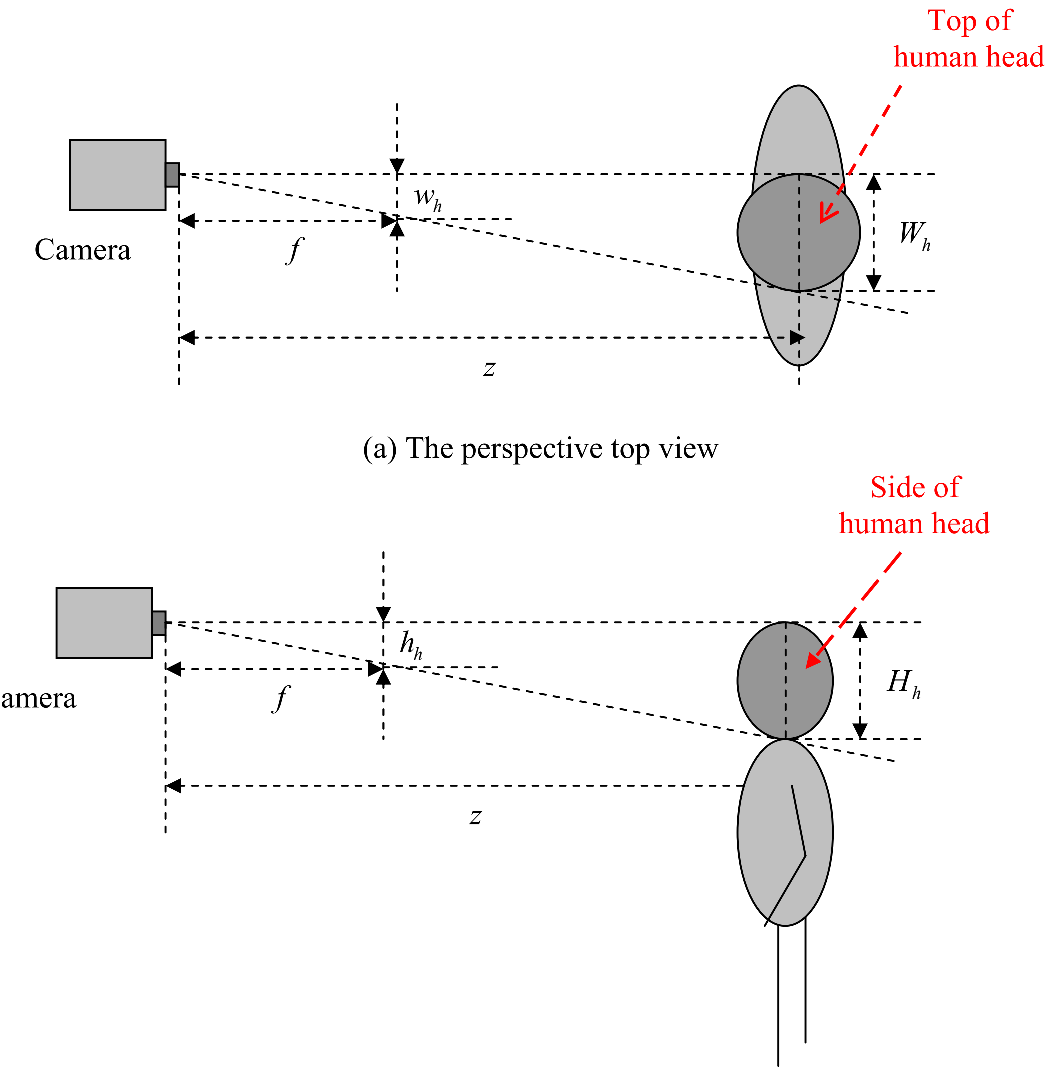
where $K_2 = fK_1$ is a constant. W_h and H_h are the average width and height of human heads, respectively.



(a) The relations between d and z: $d = K_1 / z$

(b) Distributions of human heads in disparity measure

Figure 4.4 Human head beings in different depths to cameras



(a) The perspective top view

(b) The perspective side view

Figure 4.5 The geometrical representation of the positions of camera and human head

Now we know that if a point (x, y) is a center of a human head, the cloud of the disparity data from this head will be distributed within a 3D bounding box centered at it with width, height, and depth range being w_h, h_h and $[d_-, d_+]$, respectively. Besides, since the human heads look like the isolated balls from an elevated vantage point, there is no object on the top and both sides of it with the same depth distance to the camera. This observation could be used to suppress the disparity measures for other objects, such as the parts of human bodies and background objects. From the above analysis, it can be seen that the measures of a human head with a certain distance to the camera would distribute in the corresponding ranges in both spatial and depth dimensions. Based on this observation, we propose a novel adaptive filtering method to extract and accumulate the evidence for each human head in the stereo image. We call it objective-oriented scale-adaptive filtering which always tries to apply the filter with the most suitable scale to aggregate the evidence for the possible targets with varying scales in the image.

From the estimated 3D sizes of the head, the filters for scale-adaptive filtering can be designed as follows. First, a 2D window is generated for a possible head centered at (x, y) as shown in Figure 4.6. The sizes of the window are chosen as

$$W_w^{(\gamma)} = 2\gamma w_h(d) \tag{4. 4}$$

$$H_w^{(\gamma)} = 1.5\gamma h_h(d) \tag{4. 5}$$

where γ is a scale factor to adapt to the scale variations for different persons. Here $\gamma \in \{0.8, 1, 1.2\}$ are used. With the layout as shown in Figure 4.6, the possible pixels belonging to the head would be within an ellipse approximating the head contour. Let (x', y') be a point within the window. The distance from the point to the head center can be defined as

$$d_\gamma(x', y') = \frac{(x'-x)^2}{a_\gamma^{\ 2}} + \frac{(y'-y)^2}{b_\gamma^{\ 2}} \tag{4. 6}$$

where $a_\gamma = \frac{\gamma w_h(d)}{2}$ and $b_\gamma = \frac{\gamma h_h(d)}{2}$. Let $\mathbf{u}' = (x', y')$ and $\mathbf{u} = (x, y)$ be the vectors of the positions, then a 2D spatial filter of a suitable scale with respect to *d* is defined as

$$F_S^{(\gamma)}(\mathbf{u}' - \mathbf{u}) = \begin{cases} 1, & d_\gamma(\mathbf{u}') \in [0, 0.7] \\ 2(1 - d_\gamma / 0.6), & d_\gamma(\mathbf{u}') \in (0.7, 1.3] \\ -1, & d_\gamma(\mathbf{u}') > 1.3 \end{cases} \tag{4. 7}$$

In practice, $F_S^{(\gamma)}(\mathbf{u}' - \mathbf{u})$ is a weight mask of a suitable size, which would give positive supports for the evidence within the ellipse and negative supports for that out of the ellipse. The spatial filter $F_S^{(\gamma)}(\mathbf{u}' - \mathbf{u})$ is shown in Figure 4.6. Because we are only interested in the points in the window whose disparity values are within $[d_-, d_+]$ for the possible head centered at (x, y), the depth filter of a suitable scale with respect to *d* centered at (x, y) is defined as

$$F_D(d'-d)=\begin{cases}[g(x',y')-d_-]/(d-d_-), & g(x',y')\in[d_-,d)\\ [d_+-g(x',y')]/(d_+-d,), & g(x',y')\in[d,d_+]\\ 0, & \mathit{otherwise}\end{cases} \tag{4. 8}$$

Where $d'=g(x',y')$ and $d=g(x,y)$. $F_D(d'-d)$ gives a larger weight for a disparity value closer to *d*.

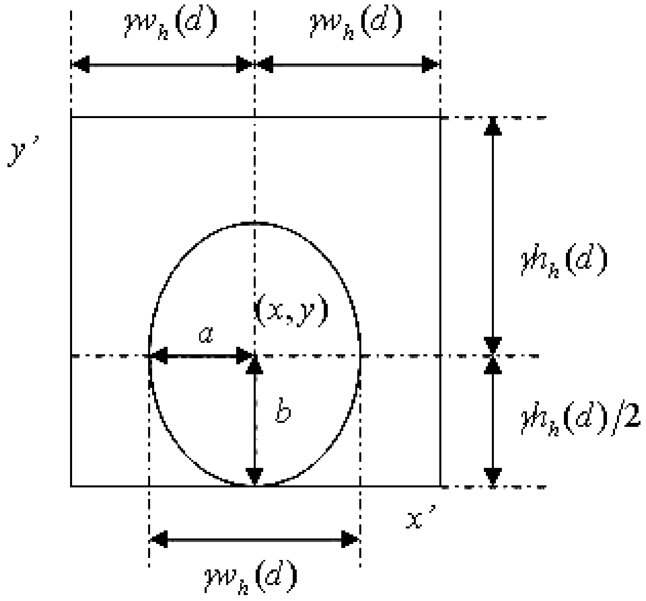


Figure 4.6 Spatial layout of the scale-adaptive filter.

The shapes of two depth filters ${F_S}^{(\gamma)}(\mathbf{u}'-\mathbf{u})$ corresponding to two different *d* values are shown in Figure 4.7.

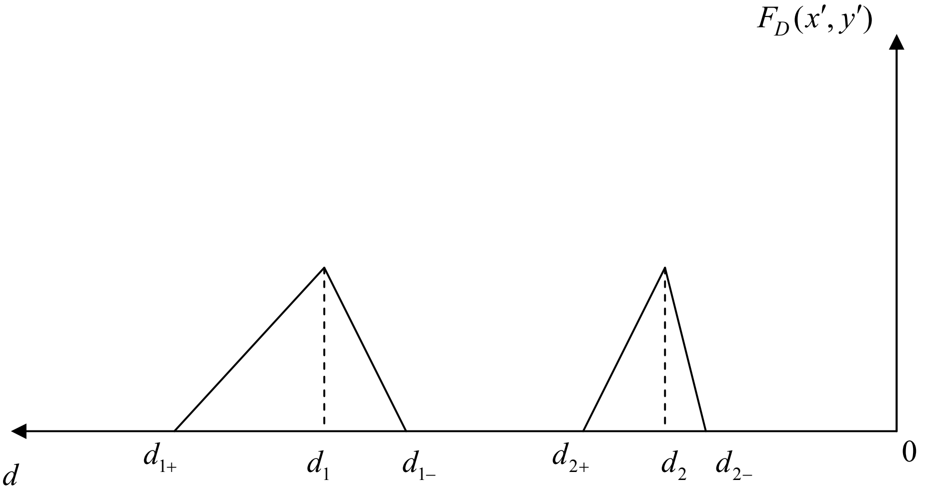


Figure 4.7 Scale-adaptive filter ${F_S}^{(\gamma)}(\mathbf{u}'-\mathbf{u})$ for human heads in different depths

With the spatial and depth filters for the pixel (x,y), the operation of scale-adaptive filtering with the scale factor γ is defined as

$$e_\gamma(x,y)=e_\gamma(\mathbf{u})=\frac{\alpha}{W_w^{(\gamma)}H_w^{(\gamma)}}\sum_{(\mathbf{u}')}{F_S}^{(\gamma)}(\mathbf{u}'-\mathbf{u})F_D(g(\mathbf{u}')-g(\mathbf{u})) \tag{4. 9}$$

where $\mathbf{u}'=(x',y')$ are the pixels within the window centered at $\mathbf{u}=(x,y)$, $g(\mathbf{u})$ is the input disparity image, and α is a constant for normalization. For *α*=3.3, the output is within

$[0, 1]$. From (4. 7), (4. 8) and (4. 9), it can be seen that, for the pixels (x', y') whose disparity values are within $[d_-, d_+]$, if they are within the ellipse they will give positive support for (x, y) belonging to a head, otherwise, they will give negative evidence of (x, y) being a part of a head. To adapt to the scale variations, the final likelihood generated by scale-adaptive filtering is defined as

$$e(x, y) = \max_{\gamma}\{e_{\gamma}(x, y)\}\text{ , } \gamma \in \{0.8, 1, 1.2\} \tag{4. 10}$$

In Equation (4.10), γ is a scale factor to adapt to the scale variations for different persons. In the likelihood map $e(x, y)$, there are several bright blobs, which correspond to possible heads. The example of the image $e(x, y)$ corresponding to Figure 3 is shown in Figure 4.8.

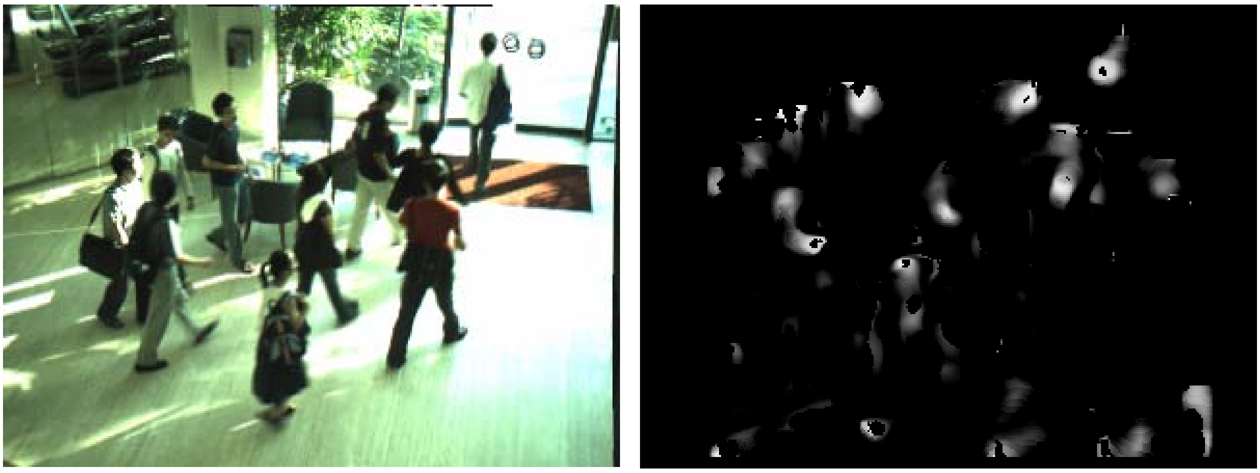


Figure 4.8 The likelihood map $e(x, y)$ of Figure 3

4.2 Suppression Of Spurious Clues

In the likelihood map $e(x, y)$, there would be some bright blobs not for human heads. They might be generated by human body parts, background objects, and shadows on the ground surface. To filter out such spurious evidence for human heads, a virtual plane, which is parallel to and above the ground surface with average human height, is established. The points in $e(x, y)$ are compared with the plane in real space and those that are far away to the plane are suppressed. The details are described as follows.

In an up-right view from an elevated vantage point, if there are two rectangles of the same size standing on the ground surface with different depth distances to the camera, the perspective geometry can be illustrated by Figure 4.9. In this case, the lines connecting corresponding corners of the two rectangles will meet at a horizontal line, the vanishing line. Let y_1 and y_2 be the vertical positions of the tops, w_1 and w_2 be the widths of the two rectangles in the image, and y_0 be the vertical position of the vanishing line. Then, from the perspective geometry and Equation (4. 1), we have

$$\frac{y_2 - y_0}{y_1 - y_0} = \frac{w_2}{w_1} = \frac{d_2}{d_1} \tag{4. 11}$$

where d_1 and d_2 are the corresponding disparity values from the two top-left corners.

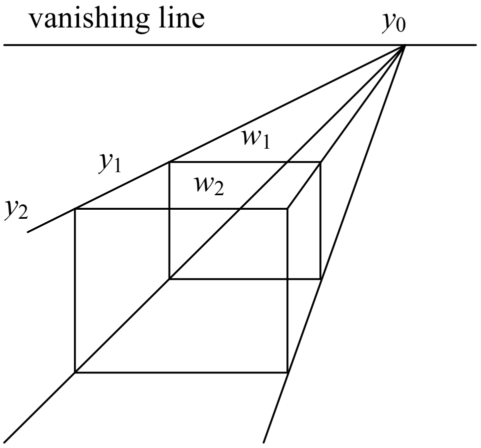


Figure 4.9 The perspective of a view from an elevated vantage point.

Let's suppose that the height of the rectangles is the average human height H_p, y^* is the position of a rectangle's top-left corner, and d^* is the disparity value of it. Then, if there is another rectangle with y and d for its top left corner from the disparity image, y can be solved, which gives a virtual plane as

$$y = Ad + B \tag{4. 12}$$

where $A = \frac{y^* - y_0}{d^*}$ and $B = y_0$ are two constants. This virtual plane is parallel to and above the ground surface with the height of H_p. It can be established in an initialization step. Let's capture several sample images with different persons in different positions in the scene, and manually mark the head positions in the images. Then we can get a set of training samples. By using Least-Square fitting, we can obtain the estimated plane

$$y = \hat{A}d + \hat{B} \tag{4. 13}$$

In the process of head detection, each pixel of $e(x, y)$ is scanned. If $e(x, y) >$ 0, from Equation (4. 3) and Equation (4. 4), the distance of the point in the real space to the virtual plane is calculated as

$$\Delta H(x, y) = \frac{| y - (\hat{A}g(x, y) + \hat{B}) |}{K_2 g(x, y)} \tag{4. 14}$$

A weight of $g(x, y)$ belonging to a head is then generated as

$$r_h(x, y) = \begin{cases} 1, & \mathit{if}\, \Delta H(x, y) \le 0.15m \\ \frac{6-20*\Delta H(x,y)}{3}, & \mathit{elseif}\, \Delta H(x, y) \le 0.3m \\ 0, & \mathit{otherwise} \end{cases} \tag{4. 15}$$

Now the suppression of the spurious evidence is performed as

$$\tilde{e}(x,y) = r_h(x,y)e(x,y) \tag{4. 16}$$

The example of the final likelihood map $\tilde{e}(x,y)$ for Figure 4.8 is shown in Figure 4.9. It can be seen that most of the spurious blobs have been suppressed and just the significant blobs for real human heads are remained.

4.3 Head Location

The last step for head detection is locating the bright blobs in $\tilde{e}(x,y)$. The heads are extracted one by one. For each human head, the following operations are performed:

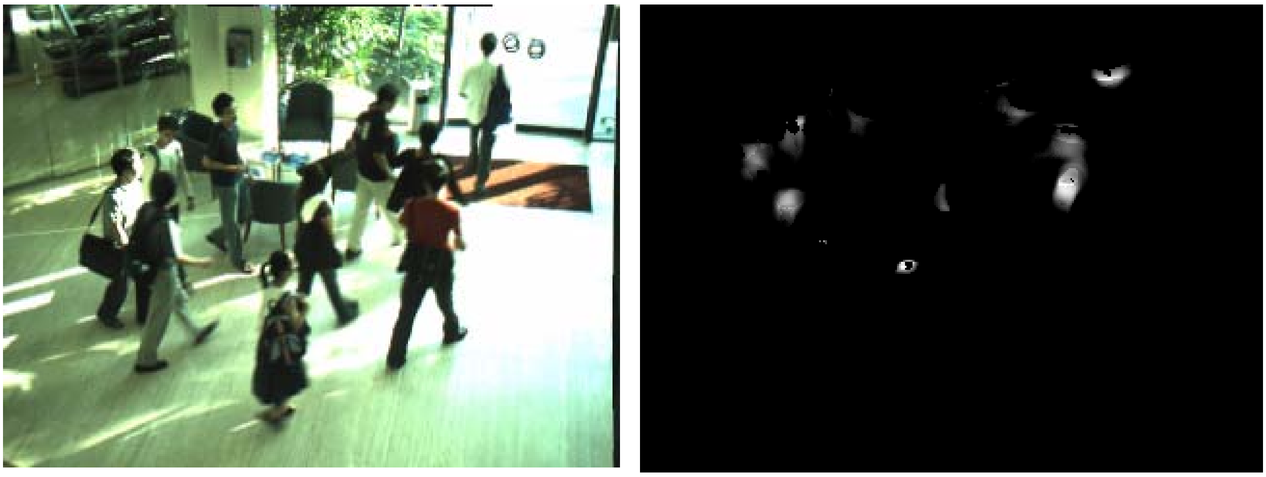


Figure 4.9 The final likelihood map $\tilde{e}(x,y)$ of $e(x,y)$ Figure 4.8

First, scan the map $\tilde{e}(x,y)$. If the maximum value $\tilde{e}(x,y)$>0.2, it is set as the initial seed to locate the head. Let $d = g(x,y)$, we can obtain $w_h(d)$ and $h_h(d)$ by Equation (4. 3). An initial window B_0 with size of $w_h(d)\times h_h(d)$ and centered at $(x_0, y_0) = (x,y)$ is established.

Secondly, a mean-shift algorithm [4] is applied to locate the head iteratively. At each iteration, the new center of the head is calculated as

$$s_t = \frac{\sum_{(x,y)\in B_{t-1}} s\tilde{e}(x,y)}{\sum_{(x,y)\in B_{t-1}} \tilde{e}(x,y)}\text{, } s = x \text{ or } y \tag{4. 17}$$

Now the window center is moved from (x_{t-1}, y_{t-1}) to (x_t, y_t). If $|x_t - x_{t-1}| < 3$ and $|y_t - y_{t-1}| < 3$ or *t*>10, the mean-shift algorithm is terminated. The position (x_t, y_t) is the detected center and B_t is the bounding box of the head. If the evidence from the window is too small, the blob is eliminated.

Finally, set $\tilde{e}(x,y) = 0$ for the pixels within the window B_t. Then start to find next human heads in $\tilde{e}(x,y)$. If no $\tilde{e}(x,y) > 0.2$ is found in a scan, the process of head detection is finished. For the detected blobs, if one is under another of larger evidence and the horizontal distance between them is less than 0.2m, it might be the shoulder and suppressed further. The remainders are the detected heads. The example of the head detection from Figure 4.1 is shown in Figure 4.10 where the detected heads are bounded with blue windows and the centers are marked with red points.

4.4 Experimental Results for Human Head Detection

The proposed stereo-based method for human head detection has been tested on the real images captured in an entrance hall of an office building from different viewpoints. The test images contain various scenes of crowds, including groups of people walking together or standing and talking to each other. By now, there is no standard database with respect to our case since not much research has been done in crowd monitoring based on stereo. Hence we can only test our method on data captured by ourselves.

To systematically evaluate the performance of the method, the results from 472 images (including those with poor illuminations) are compared with ground truth by hand. The evaluation results are summarized in Table 4.1, where $N_p = 1819$ is the number of valid persons (ground truth); $N_d = 1711$ is the number of detected persons; $N_f = 145$ is the number of the false alarms (non-human detected). The *Detection rate* indicates how many truth persons are detected with respect to the ground truth, and the *False alarm rate* shows percentage of false detections in the total detected persons. The statistics show that promising results have been achieved by the proposed method.

	Number	**Rate**
Correct Detections	$N_d = 1711$	$\frac{N_d}{N_p} = 94.06\%$
False Alarms	$N_f = 145$	$\frac{N_f}{N_d + N_f} = 7.81\%$

Table 4.1 Systematic Evaluation for Head Detection

Some examples are illustrated in the rest of this section. One example of 10 persons walking together is shown in Figure 4.10. In the example images, there are some misalignments of head locations. This is because the disparity images are generated after warp rectification and by a region-based method [15]. Another example of 6 persons walking in the hall is shown in Figure 4.11. One more example of 6 persons standing and talking to each other is shown in Figure 4.12. In these two examples, there are many cases of two persons overlapping in the depth direction, which are the most difficult cases for the methods on the 2D regions [28]. The proposed method detects the persons correctly for these tough cases. More examples are illustrated in Figure 4.13~16. Some edge area doesn't have image in both the left and the right camera, so the correlation cannot be conducted. Hence there is no stereo information in this kind of area. That is the reason why the detection result for the man in the right edge area leans a bit to the center direction in Figure 4.13, while there is no disparity information for the guy in the left edge in Figure 4.15. In some sequences, we confine the scene area within a certain boundary. For example, the area beyond the gate in Figure 4.14 is not concerned. So some people near or beyond the gate are not detected. As illustrated in Figure 4.16, the images were taken at night when the lighting condition was not good. Besides, the reflections of the lights on the ground are confusing. Our method shows it is robust even in poor illumination condition.

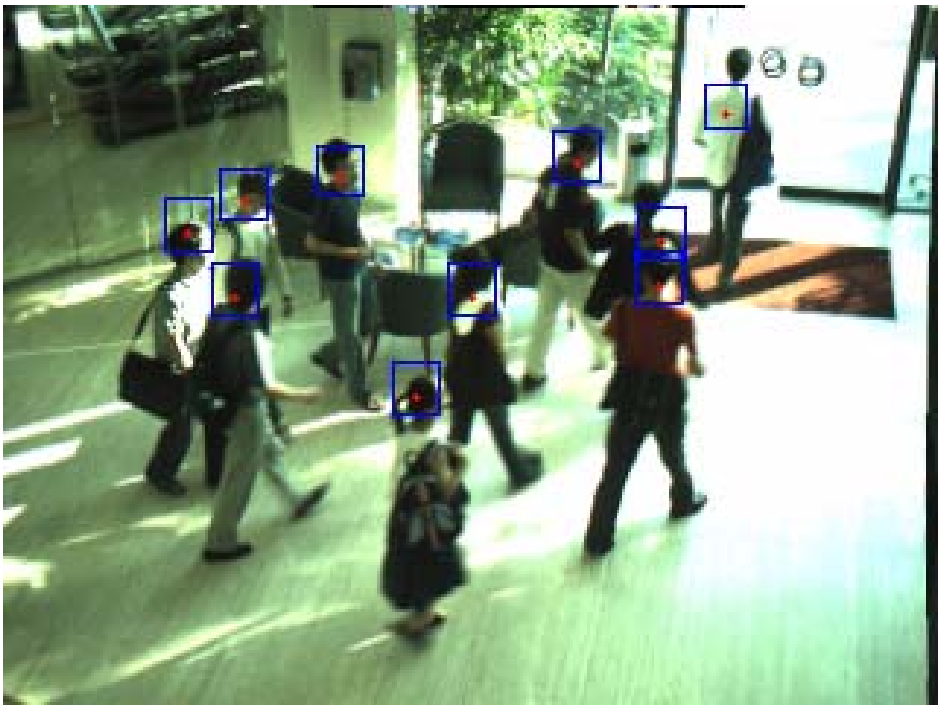
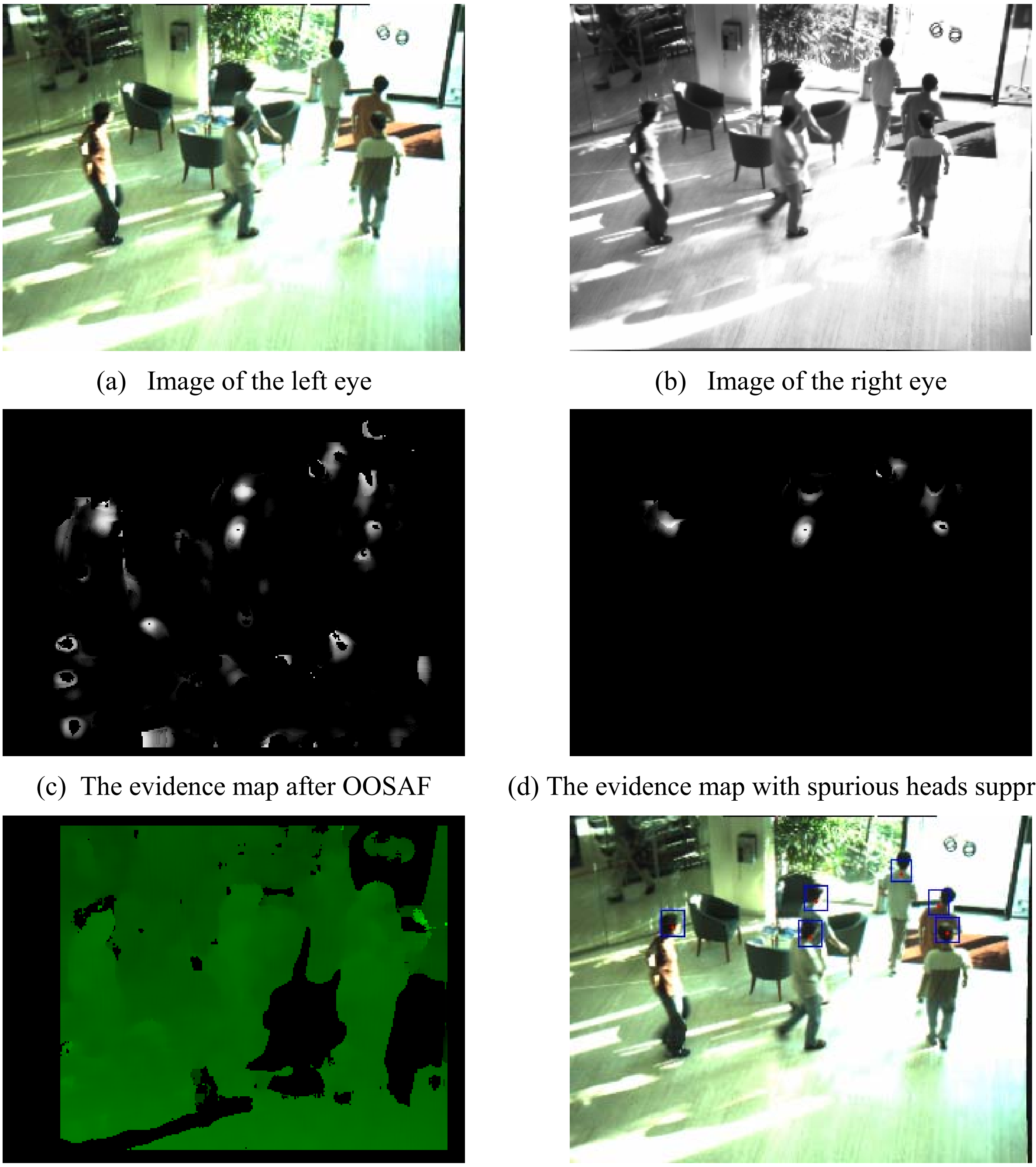


Figure 4.10 The final result of head detection from Figure 4.1



(a) Image of the left eye

(b) Image of the right eye

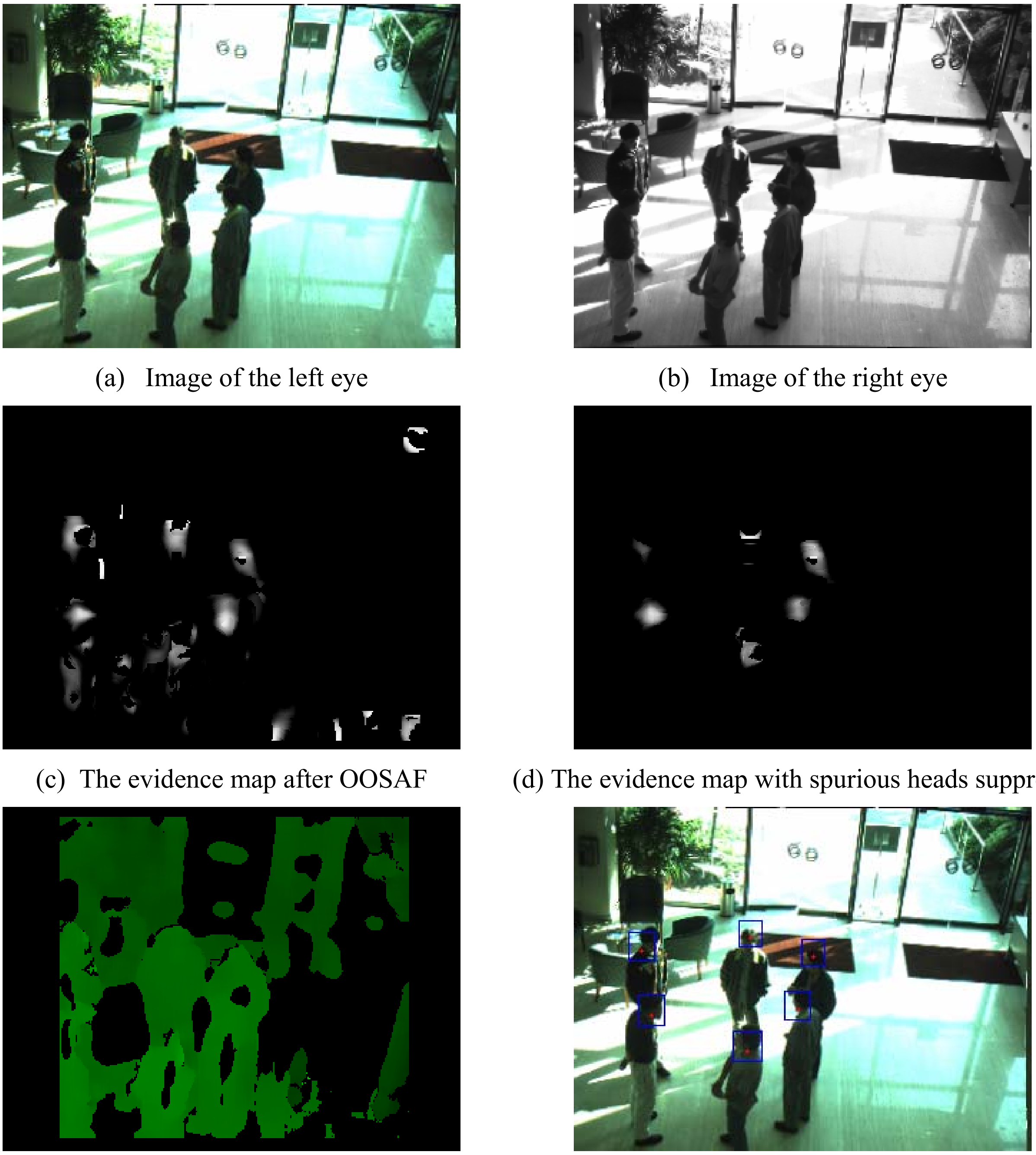
(c) The evidence map after OOSAF

(d) The evidence map with spurious heads suppressed

(e) The disparity image

(f) The final detection result on the color image

Figure 4.11 An example of detection results of walking group made up of 6 people



(a) Image of the left eye

(b) Image of the right eye

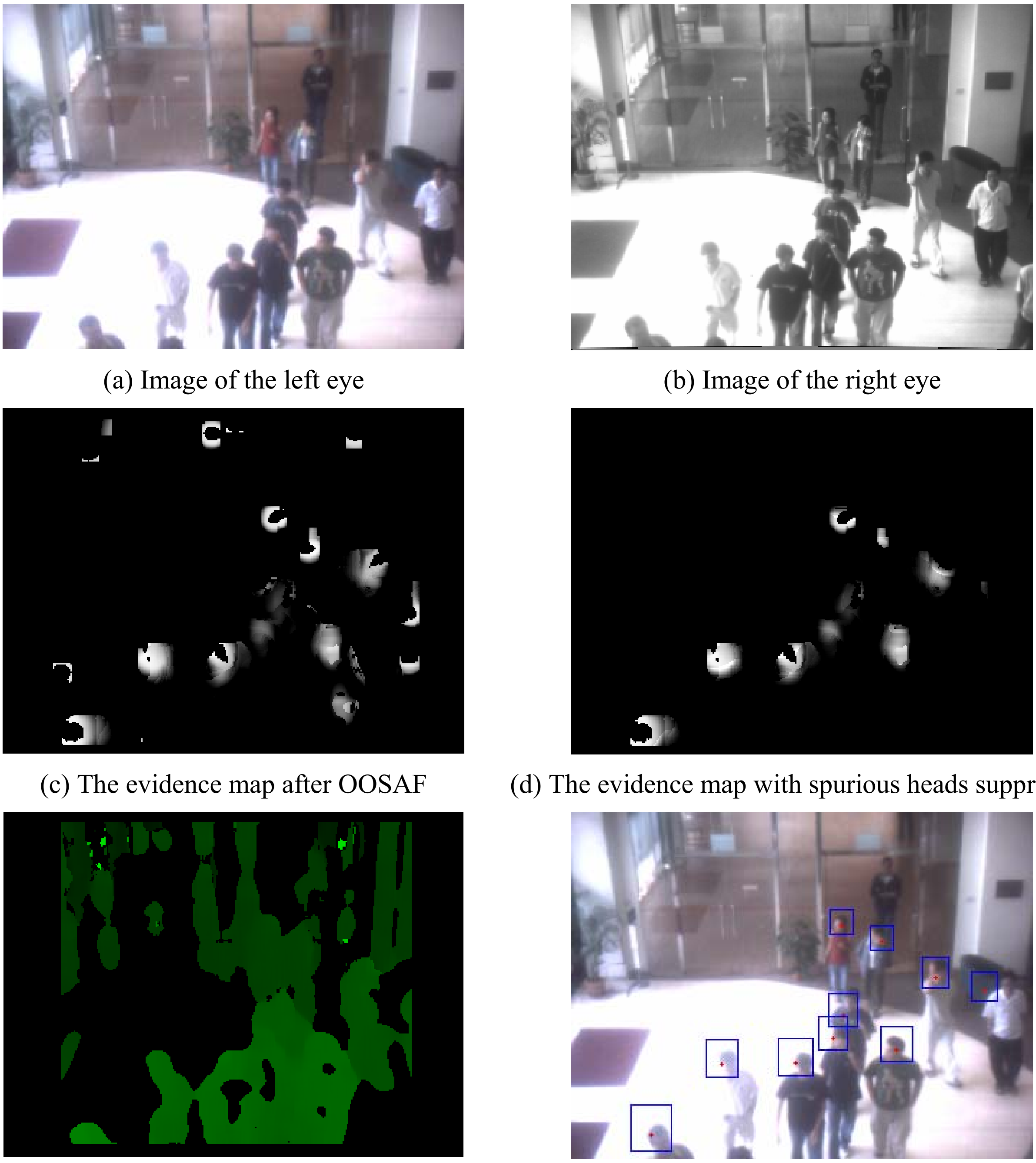
(c) The evidence map after OOSAF

(d) The evidence map with spurious heads suppressed

(e) The disparity image

(f) The final detection result on the color image

Figure 4.12 An example of detection results of talking group made up of 6 people



(a) Image of the left eye

(b) Image of the right eye

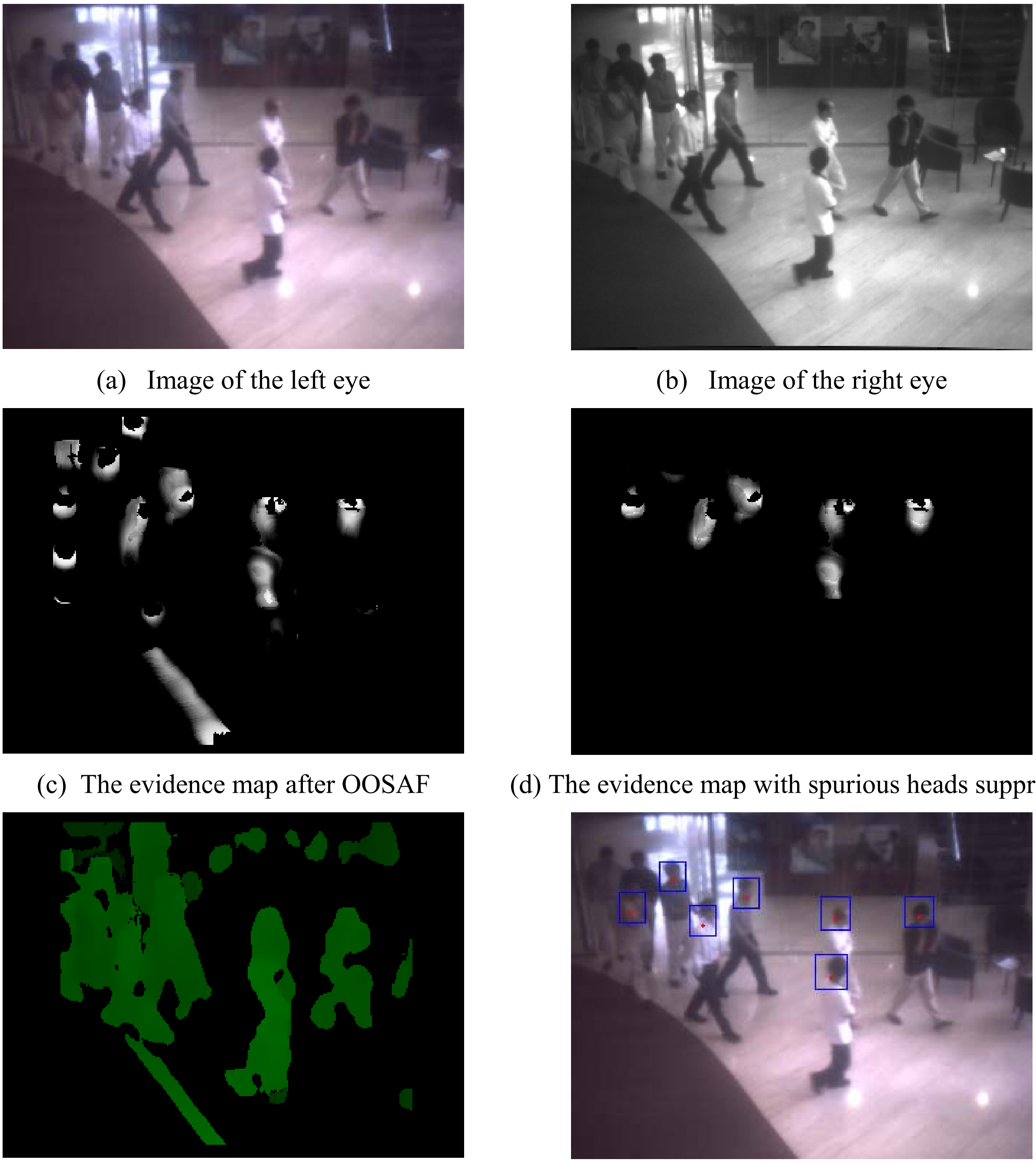
(c) The evidence map after OOSAF

(d) The evidence map with spurious heads suppressed

(e) The disparity image

(f) The final detection result on the color image

Figure 4.13 More examples: walking group made up of 10 people



(a) Image of the left eye

(b) Image of the right eye

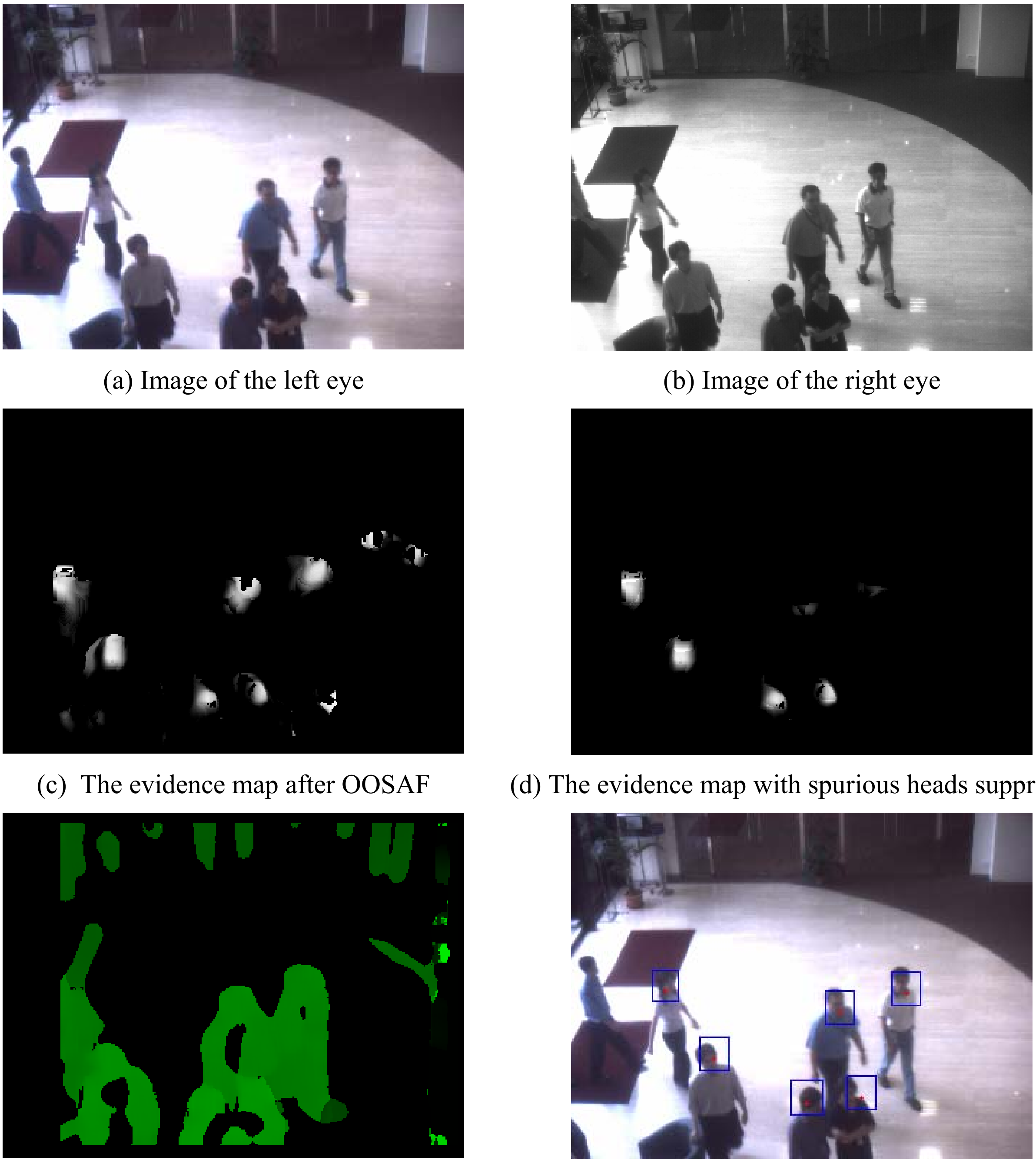
(c) The evidence map after OOSAF

(d) The evidence map with spurious heads suppressed

(e) The disparity image

(f) The final detection result on the color image

Figure 4.14 More examples: walking group made up of 7 people



(a) Image of the left eye

(b) Image of the right eye

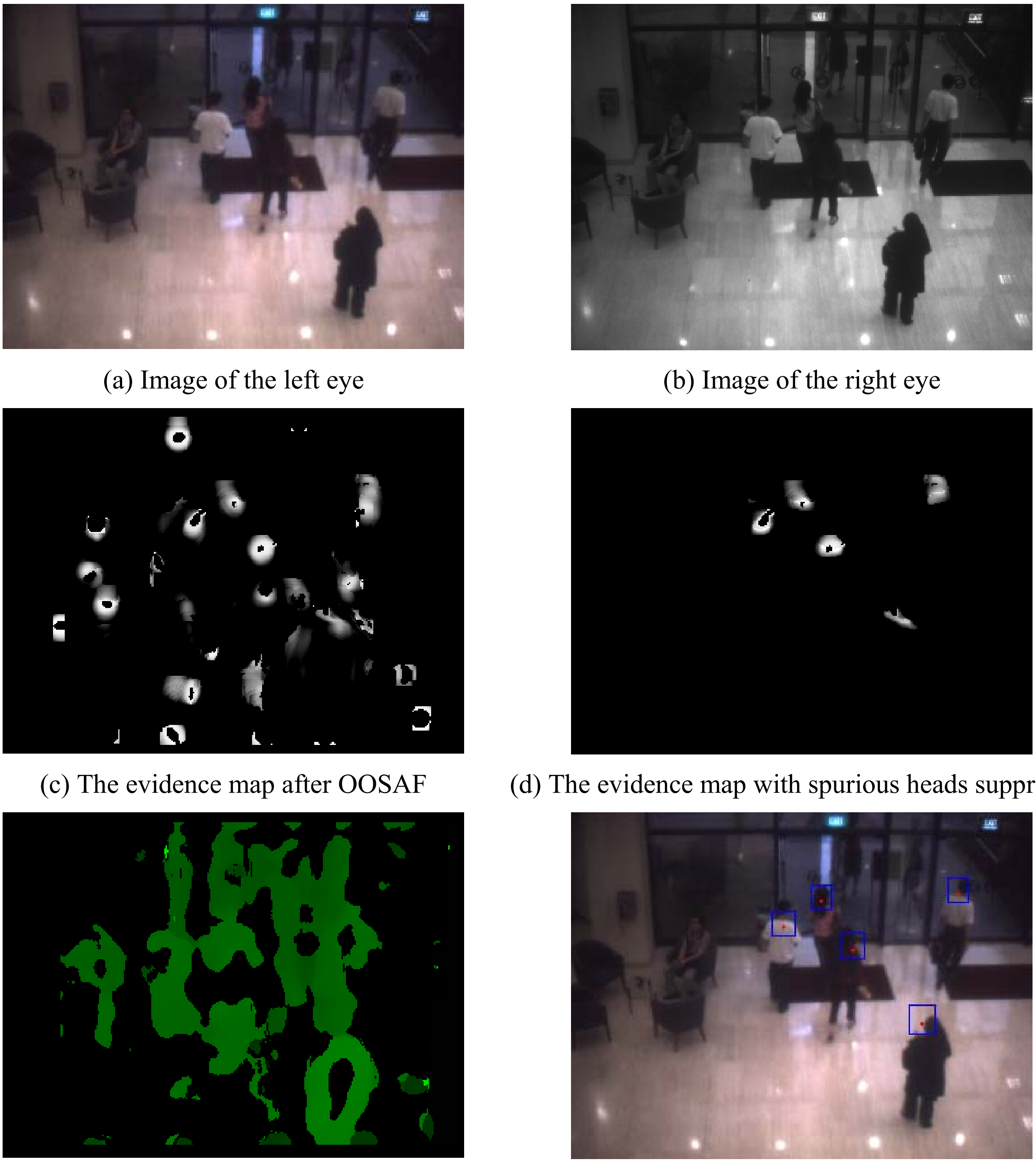
(c) The evidence map after OOSAF

(d) The evidence map with spurious heads suppressed

(e) The disparity image

(f) The final detection result on the color image

Figure 4.15 More examples: walking group made up of 6 people



(a) Image of the left eye

(b) Image of the right eye

(c) The evidence map after OOSAF

(d) The evidence map with spurious heads suppressed

(e) The disparity image

(f) The final detection result on the color image

Figure 4.16 More examples: crowd of 5 people with dark illumination

From the examples, we can see our method for human head detection works well even in very crowded scene. Also the method proves itself to be robust even in sequence with poor illumination. The results demonstrate the robustness of our head detection method.

Chapter 5

Head tracking

With the good results of human head detection in each frame, we try to track each head for the spatial-temporal description of the motion for each person, such as the motion direction, speed, and trajectory. The information is very important for crowd monitoring. Individuals can be clustered into groups based on the motion information, and then group behavior analysis may be conducted further. Besides, tracking with the motion information derived from the consecutive frames can help to rectify some errors in head detection that depends on a single frame only.

When a human object is moving in the scene, the image of his/her head as well as the surrounding region keeps changing. In some frames, the head might be unclear in the color images, and in some other frames, the stereo measures of the head could be incomplete. Hence, fusing the information of motion, color and stereo may lead to better tracking result. In the rest of this chapter, the proposed tracking method is described in two parts: motion prediction and data association (matching).

5.1 Prediction with Kalman filter

In our system, Kalman filter is employed for motion prediction. In this case, human heads are seldom occluded if the camera is located at an elevated position. When the frame rate is high enough, the motion of the human head through a few consecutive frames can be described by linear dynamic models. Hence the Kalman filter is accurate enough to track each human head through the frames.

With the stereo images, we can track the human heads in 3D space. The center position of a head directly from the stereo image is $\hat{\mathbf{p}} = (x, y, d)^T$. Since the relation between the disparity *d* and the depth distance *z* is nonlinear (See Equation (4. 1)), when a person $\mathbf{p} = (x, y, z)^T$ is used for Kalman filter, *z* is converted from *d*.

$$\begin{cases} \mathbf{x}_i = \mathbf{D}\mathbf{x}_{i-1} + \omega \\ \mathbf{y}_i = \mathbf{M}\mathbf{x}_i + \varepsilon \end{cases} \tag{5. 1}$$

where ω and ε are Gaussian noises. $\mathbf{D}$ is the dynamic matrix and $\mathbf{M}$ is the measurement matrix. Let $\mathbf{p} = (x, y, z)^T$ be the position vector, $\mathbf{v} = (\dot{x}, \dot{y}, \dot{z})^T$ be the vector of velocity, and $\mathbf{a} = (\ddot{x}, \ddot{y}, \ddot{z})^T$ be the vector of acceleration. Assume that the human head moves with a constant acceleration during a short period of time. In this case, $\mathbf{p}_i = \mathbf{p}_{i-1} + (\vartriangle t)\mathbf{v}_{i-1}$,

$\mathbf{v}_i = \mathbf{v}_{i-1} + (\triangle t)\mathbf{a}_{i-1}$, and $\mathbf{a} = \mathbf{a}_{i-1}$. Hence, the state vector for Kalman filter is $\mathbf{x} = [\mathbf{p}, \mathbf{v}, \mathbf{a}]^T$ and the vector of the observed measurements is $\mathbf{y} = [\mathbf{p}, 0, 0]^T$. The dynamic matrix $\mathbf{D}$ and the measurement matrix $\mathbf{M}$ in Kalman filter become

$$\mathbf{D} = \begin{bmatrix} \mathbf{I} & (\triangle t)\mathbf{I} & 0 \\ 0 & \mathbf{I} & (\triangle t)\mathbf{I} \\ 0 & 0 & \mathbf{I} \end{bmatrix} \tag{5. 2}$$

$$\mathbf{M} = \begin{bmatrix} \mathbf{I} & 0 & 0 \end{bmatrix} \tag{5. 3}$$

The Kalman filter is initialized with the head positions in the first three frames when the head is detected and tracked directly. Suppose a human head is detected firstly at time t_0. Then, with $\mathbf{p}_0$, $\mathbf{p}_1$, and $\mathbf{p}_2$, the speed vector is calculated as

$$\mathbf{v}_1 = \frac{\mathbf{p}_1 - \mathbf{p}_0}{\triangle t}, \mathbf{v}_2 = \frac{\mathbf{p}_2 - \mathbf{p}_1}{\triangle t} \tag{5. 4}$$

and the vector of accelerations is

$$\mathbf{a}_2 = \frac{\mathbf{v}_2 - \mathbf{v}_1}{\triangle t} \tag{5. 5}$$

Now the initial state vector becomes

$$\mathbf{x}_2 = \begin{bmatrix} \mathbf{p}_2 \\ \dfrac{\mathbf{v}_1 + \mathbf{v}_0}{2} \\ \mathbf{a}_2 \end{bmatrix} \tag{5. 6}$$

Afterwards, we can predicate the position $\hat{\mathbf{p}}_i$ from the previous observations $\mathbf{p}_{i-1}$ with the Kalman Filter. Here, with a constant frame rate, $\vartriangle t$ can be set as 1.

5.2 Data association

Data association is to find a new position of the tracked object with the best observation measure. In this case, to find the new position is to search for the head location around the predicted position in the current frame, which actually is to find the area that gives the best matching to the previous head appearance. Stereo images can provide good information for isolated objects in 3D space in many cases. On the other hand, visual information, such as color information, is very informative for object appearance in image sequences. In some places in the image, both stereo and color information might not be good enough due to bad imaging conditions, such as bad illumination. The motion information provided by Kalman filter can compensate the deficiency of the above two kinds of information.

The main difficulty for tracking method based on multi-feature is to integrate the multiple cues into one framework, and adapt them to the changing appearance of the objects and the background. Here we propose an adaptive method to fuse the information of motion, color, and stereo based on kernel function since the kernel functions is robust to the variation of the measures.

Let $(\hat{x}_i, \hat{y}_i)$ be the predicted head center in the current frame produced by Kalman filter, and (x, y) be a neighbor point of $(\hat{x}_i, \hat{y}_i)$ in the current frame. The point (x, y) can be a candidate for the new position of the head being tracked. Assume (u, v) be a point within a window centered at (x, y) where the origin of the coordinate is (x, y), e.g. $u \in [-W, +W]$, $v \in [-H, +H]$. $W \times H$ is the size of the window for the head being tracked. Now, the measures for head matching can be evaluated as follows:

The Spatial kernel: Human heads look like circle objects with certain sizes in the 2D image. Hence, a scale-adaptive spatial kernel $k_s(u, v)$ is designed to weigh the evidence for the human head in the window centered at (x, y). The spatial kernel is defined as:

$$k_s(u,v) = \exp\left\{ -\frac{\| (u,v) \|^2}{s^2(\hat{d}_i)} \right\} \tag{5. 7}$$

where $s(\hat{d}_i) = K_2 W_h \hat{d}_i$ and $\hat{d}_i = K_1 / \hat{z}_i$. W_h is the average width of human heads and K_2 is a constant as described in Chapter 4.

The spatial kernel enhances the information within the circle centered at (x, y) with a radius corresponding to the width of the human head in the image. The more close to the center of the circle, the bigger the weight is. By applying the spatial kernel, we can suppress the information outside the potential human head area and focus on the information inside it.

Color similarity: When an object moves in the scene, the colors of the object in the consecutive frames will vary in certain extent. Here, a color kernel function is designed to evaluate the similarity between the candidate head area in the current frame and the tracked head area in the previous frame. Let (x, y) be the center of the candidate window, (x_{i-1}, y_{i-1}) be the center of the head in the previous frame, and $\mathbf{u} = (u, v)$ be a point within the current candidate window respect to the center $\mathbf{x} = (x, y)$. The similarity function $k_c(\vartriangle\mathbf{u})$ is defined as:

$$k_c(\vartriangle\mathbf{u}) = \exp\left\{-\frac{|\mathbf{I}_i(\mathbf{x}-\mathbf{u}) - \mathbf{I}_{i-1}(\mathbf{x}_{i-1}-\mathbf{u})|^2}{{\sigma_c}^2}\right\} \tag{5. 8}$$

where $\triangle\mathbf{u}=\mathbf{x}-\mathbf{u}$, $\mathbf{I}_i(x,y)$ and $\mathbf{I}_{i-1}(x,y)$ are the color vector (RGB values) of the point (x,y) in the current and previous frames respectively. σ_c is a scale parameter for color variation, which should be properly chosen to make the kernel work effectively.

Depth distance: To evaluate how close in the depth dimension between the candidate head and the predicted head, a kernel function for depth distance is defined. Since the disparity measures are non-linear with respect to the distance from the object to the camera, they are converted to depth measures *z* before evaluation. The kernel function $k_d(\triangle\mathbf{u})$ is defined as follows:

$$k_d(\triangle\mathbf{u})=\exp\left\{-\frac{|Z_i(\mathbf{x}-\mathbf{u})-\hat{z}_i)|^2}{{\sigma_z}^2}\right\} \tag{5. 9}$$

where $\sigma_z=D_h/2$, and D_h is the average depth of human heads as in Equation (4.2). $z_i(x,y)=\dfrac{K_1}{g_t(x,y)}$ and $g_i(x,y)$ is the disparity value of (x,y) in the current frame. As $\hat{z}_i$ is the estimated depth for the head center in the current frame, actually, $k_d(\triangle\mathbf{u})$ is used to evaluate the similarity of the head between the previous and the current frames in the 3rd dimension *z*.

Head Existence: Meanwhile, the evidence of head existence generated from the scale-adaptive filtering can also provide a good estimation about whether there is a human head around the predicted position. The blobs in the evidence map can be used as a weight mask for head tracking. Hence, the function of head existence is defined as:

$$k_h(\triangle \mathbf{u}) = \tilde{\mathrm{e}}_i(\mathbf{x} - \mathbf{u}) \tag{5. 10}$$

where $\mathbf{x} = (x, y)$ is the center of the candidate and $\mathbf{u} = (u, v)$ is a point within the candidate window respect to (x, y). As defined in Equation (4. 16), $\tilde{e}_i(x, y)$ represents the possibility that (x, y) belongs to a human head area [12]. Assisted by the head existence function, the output of the scale-adaptive human head detection be integrated into the tracking.

Candidate Matching: With the above four kernels in hand, we need to fuse them together to determine the general evaluation of the matching result. By multiplying the weights of the four kernels, and using $\sum_{\mathbf{u}} k_s(\mathbf{u})$ to normalize the final matching weight, we can obtain the matching result with the following formula:

$$m(\mathbf{x})=\frac{\sum_{\mathbf{u}} k_s(\mathbf{x}-\mathbf{u})\cdot k_h(\mathbf{x}-\mathbf{u})\cdot k_d(\mathbf{x}-\mathbf{u})\cdot k_c(\mathbf{x}-\mathbf{u})}{\sum_{\mathbf{u}} k_s(\mathbf{x}-\mathbf{u})}\cdot w(\mathbf{x}) \tag{5. 11}$$

where the weight $w(\mathbf{x})=w(x,y)$ takes the shift of the head from the predicted position $(\hat{x}_i,\hat{y}_i)$ into account. It is defined as:

$$w(\mathbf{x})=w(x,y)=\exp\left\{-\frac{|(x,y)-(\hat{x}_i,\hat{y}_i)|^2}{s^2(\hat{d}_i)}\right\} \tag{5. 12}$$

Now, the new position of the tracked head in the current frame is determined as follows. Let

$$({x^*}_i,{y^*}_i)=\arg\max_{(x,y)}\{m(x,y)\} \tag{5. 13}$$

Then, the observation position of the head in Frame *i* is decided by:

$$(x_i,y_i)=\begin{cases}({x^*}_i,{y^*}_i), & \mathit{if} \quad m({x^*}_i,{y^*}_i)>T\\ (\hat{x}_i,\hat{y}_i), & \mathit{otherwise}\end{cases} \tag{5. 14}$$

where *T* is a threshold for the possibility of the matching.

The head location at Frame i is obtained by finding the position with the maximum $m(x, y)$ larger than a certain threshold, which means the position has the smallest and acceptable difference with respect to color, stereo, and motion information, compared to all the other candidate positions in the current image. However, if the largest $m(x, y)$ is still smaller than the threshold, which means the difference between the most possible position and the tracked head position in the previous frame is so big that the head is considered as lost in the current frame. This may result from various factors including poor illumination and bad disparity measures. For example, when a head with dark hair appears in a dark background, disparity measures may be missed in this area since there is too little texture information inside the region. If the head moves out of the dark background, the disparity measures are available again. In some situations such as the above case or sudden illumination changes, if we make the system rely on the motion information provided by the Kalman filter, hopefully the algorithm can recover the tracked head when the situation gets better after several frames.

5.3 Experimental Results of Human Head Tracking

The proposed method for human head tracking has been tested on several real sequences captured in an entrance hall of an office building as described in the previous chapter. The test sequences contain scenarios of people walking separately, two or three persons walking together, and groups of people moving as human flows in the scene. The frame rates of the test sequences are about 8 frames per second. By now, there is no standard database with respect to our case since not much research has been done in crowd monitoring based on stereo. Hence we can only test our method on data captured by ourselves. There are some limitations to capture good test sequences at public areas, e.g., the power supply and the permission of the authority. Hence our data for test is limited.

A systematical evaluation for the performance of the tracking method has been done on several image sequences. A comparison between our detection and tracking methods is shown in Table 5.1 on detection rate and false alarm rate. Another comparison on head shifts with respect to the ground truth is illustrated in Figure 5.4. Also, a comparison between our tracking method and a simple Kalman model with constant acceleration using color-based matching is shown in Table 5.2.

In Table 5.1, the results of our detection and tracking methods from 458 frames taken from three sequences are compared with ground truth by hand. N_d and N_f are the numbers of the correct detections and the false alarms for detection respectively; N'_d and N'_f are the numbers for tracking. N_p is the number of valid persons (ground truth). From the experiment results we can see that tracking method, which uses color information only and use stereo-based detection result for initiation, slightly improves the correct detection rate of the stereo-based detection method, which is based on stereo information in one single frame only. That is because motion information can help to continuously track the target, which is not detected by the stereo-based method only in a few frames. However, the false alarm rate of the color-based method is higher because the false alarm initiation leads to the consequent tracking. The fusion tracking method, which uses motion information as well as color information in the consecutive frames, leads to better results compared to the stereo-based detection method and the color-based tracking method.

	Sequence No. (No. of Frames)	N_p	Number			Rate		
			Stereo-based Detection	Fusion Tracking	Color-based Tracking	Stereo-based Detection	Fusion Tracking	Color-based Tracking
Correct Detections (Rate: $\frac{N_d}{N_p}$)	No. 1 (199)	263	263	263	263	100%	100%	100%
	No. 2 (199)	240	235	240	237	97.92%	100%	98.75%
	No. 3 (60)	127	113	126	113	88.98%	99.21%	88.98%
	Total (458)	630	611	629	613	96.98%	99.84%	97.30%
False Alarms (Rate: $\frac{N_f}{N_d+N_f}$)	No. 1 (199)	263	4	0	12	1.50%	0	4.36%
	No. 2 (199)	240	1	0	8	0.42%	0	3.23%
	No. 3 (60)	127	17	16	38	13.08%	11.27%	23.03%
	Total (458)	630	22	16	58	3.37%	2.48%	8.43%

Table 5.1 Systematic Evaluation for Head Tracking Compared with Stereo-Based Detection (N_p is the ground truth.)

In Sequence No. 1 made up of 194 frames that contains 3 people in total, the tracking results are all correct, while there are 4 false detected heads generated by the head detection method. In Sequence No. 2 consisting of 199 frames, which contains 2 people in total, both the tracking and detection results are all correct. In 70 frames of Sequence No. 3 containing 199 frames, two people walk side by side. One of the two persons is lost and then tracked by a different ID number for once. The same thing happens to the other one twice.

Figure 5.1 shows one example frame from a sequence in which one person is standing while two others are walking together from the left to the right. In the figure, the yellow bounding boxes indicate the tracked head areas and the numbered yellow stars mark the corresponding centers. The detection results are superimposed on the image as blue bounding boxes for comparison. The tracking results can provide space-temporal information of each person, i.e., the velocity, the direction of movement, and the trajectory. The tracking method not only provides the motion information of each individual, but also corrects some errors of the head detection results obtained from the single frame.

In Figure 5.2, there is a false detected head near the head of the upper person. This error is eliminated in the tracking result by using the motion and color information of the preceding frames as well as the current frame. A similar case can be found in Figure 5.3.

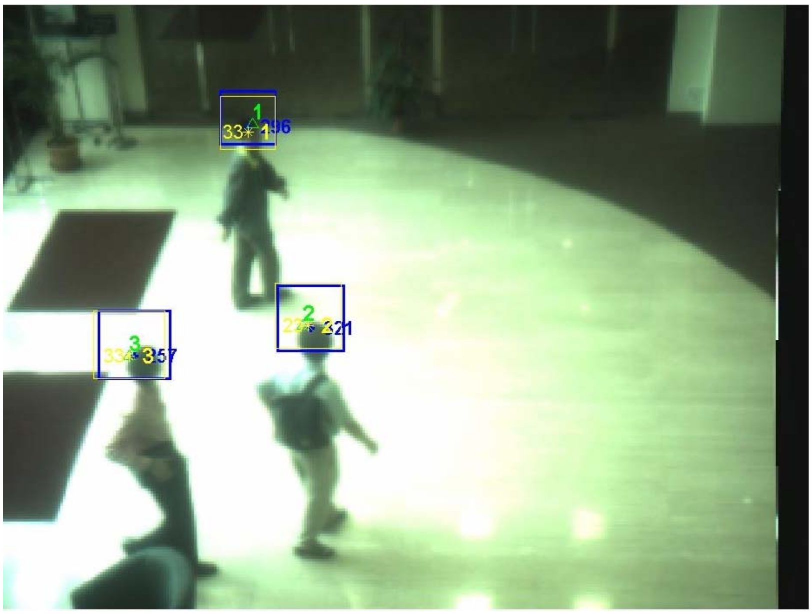


Figure 5.1 An example of tracking results (yellow) as well as detection results (blue)

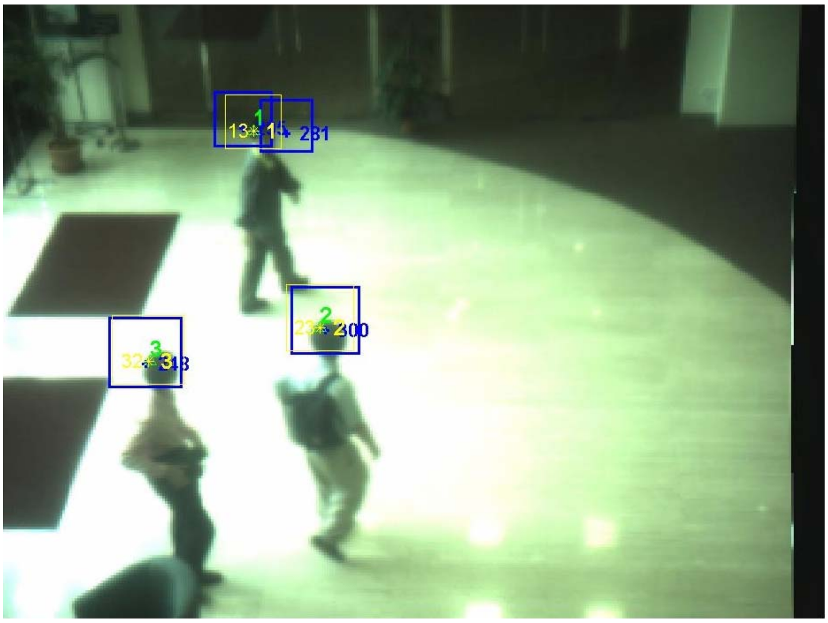


Figure 5.2 The false alarm in the detection result has been eliminated in the tracking result

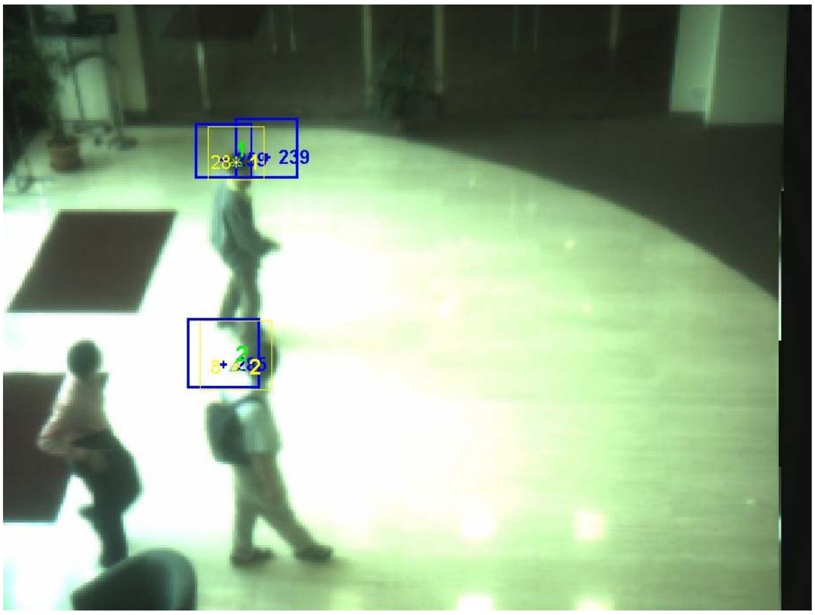


Figure 5.3 The tracking method leads to more accurate head position than the detection does (including one false alarm of detection method)

Fusing the temporal, color, and stereo information, the tracking method can generate more accurate positions of the tracked heads than the head detection does, which just use information in one single stereo image. One example is shown in Figure 5.3 where the centers of the yellow bounding boxes, which are indicated by yellow stars, are more close to the real head centers. The head shift with respect to the ground truth marked manually for both head detection and tracking has also been evaluated. One example of the comparison on the sequence shown in Figure 5.3 is demonstrated in Figure 5.4, where the blue curve indicates the head shifts of head detection results and the read one indicates the head shifts for tracking results. The average of the head shifts from the tracking results of one human head in 50 frames is 4.28 pixels, while the average of the head shift of the detection results is 5.63 pixels. The tracking method leads to a more accurate head position than the detection method, due to the employment of the multiple cues, which can be seen from the figure clearly.

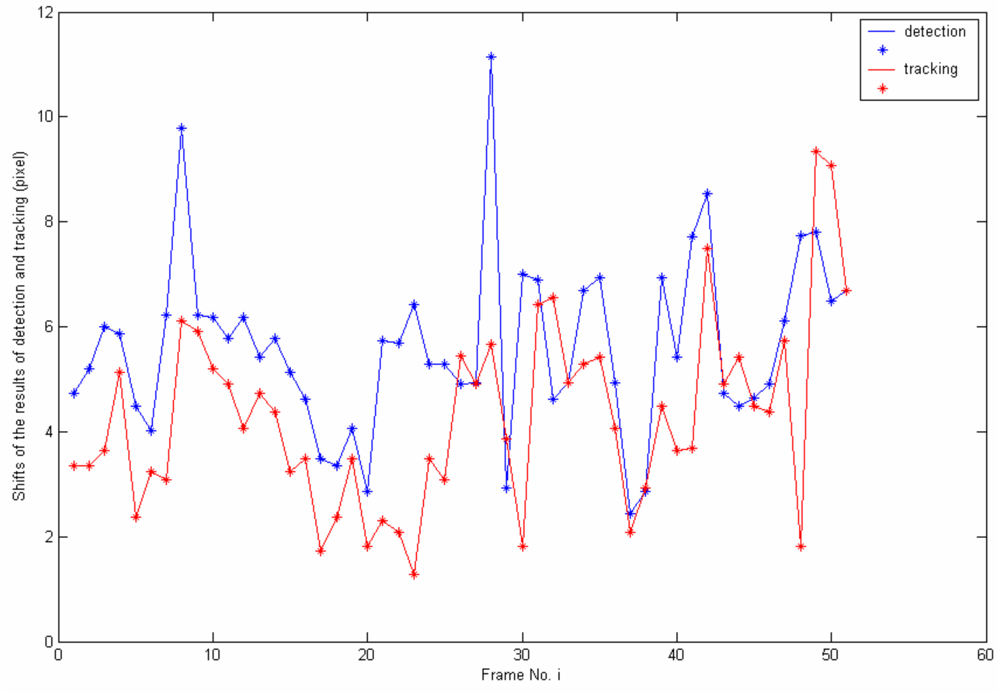
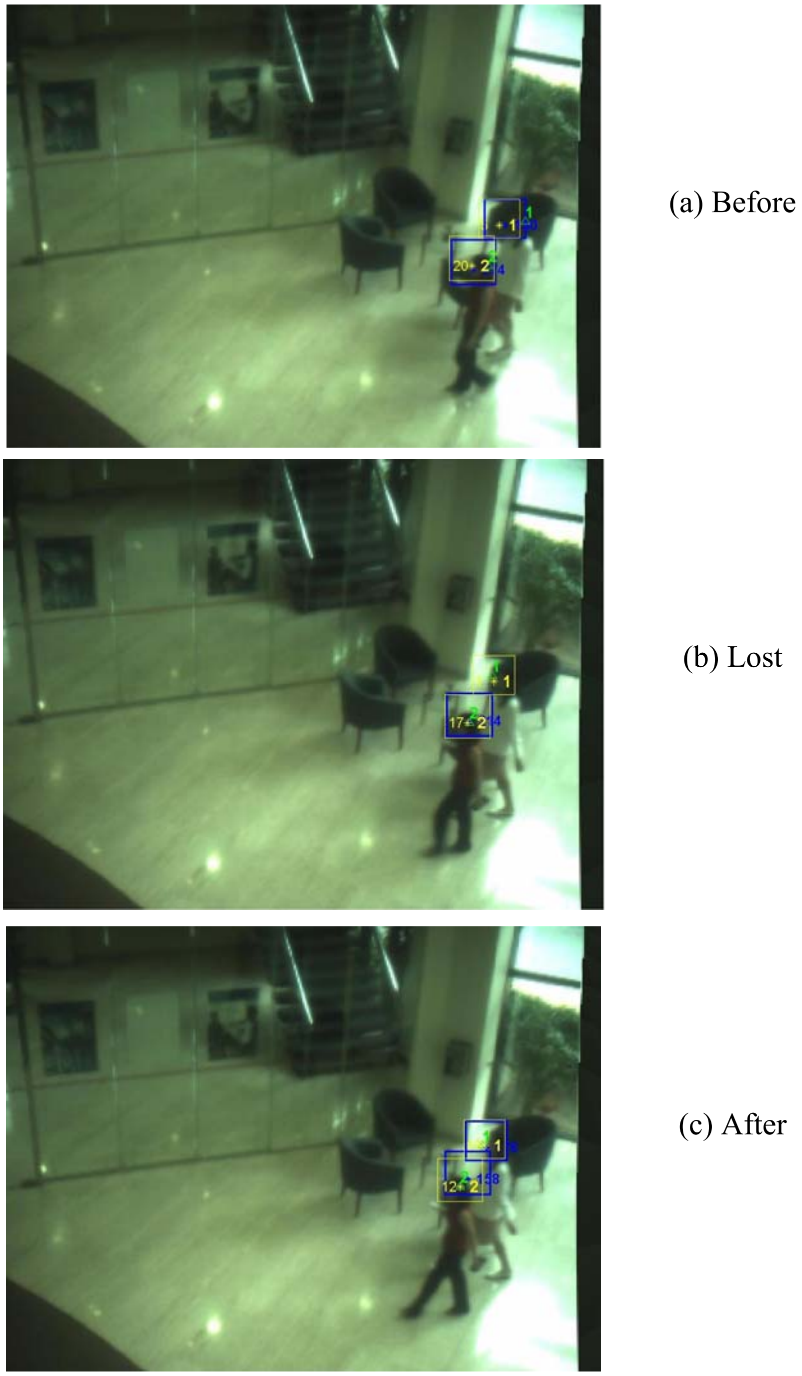


Figure 5.4 The head shift for head detection and tracking

In some situations, when the stereo information or the color information is not sufficient, the detection method may fail since it is based on the stereo information only. However, the tracking method can follow the head position on the motion information and recover it when the stereo information or color information is available again. One example of such situation is shown in Figure 5.5.



(a) Before

(b) Lost

(c) After

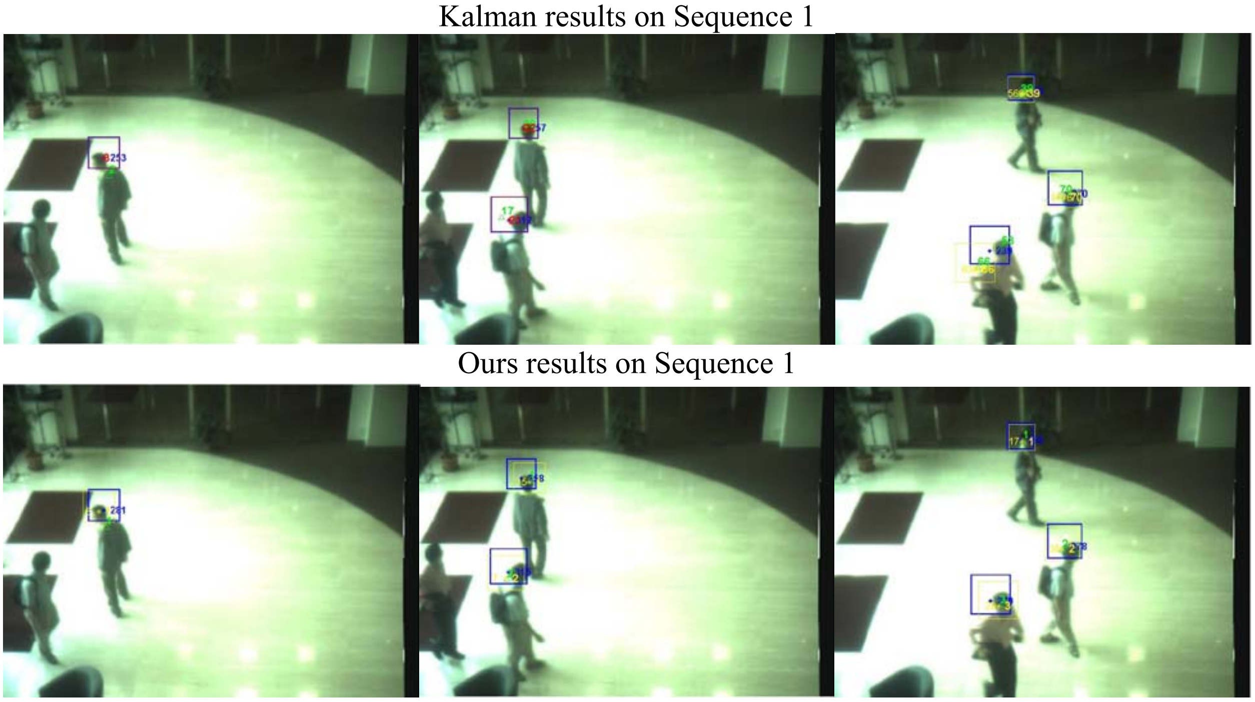
Figure 5.5 The tracking method worked well while detection lost

Figure 5.5 contains three images. In the middle image, the head of the person on the right with black hair is in front of a dark background region. The color and stereo information about the head is poor in this area. Hence, head detection method fails in this case. The first and third images are the frames just before and after the head being in front of the dark background region. It can be seen that the head was tracked successfully when the head passing through the dark background region.

A comparison between our method and a simple Kalman model with constant acceleration using color-based matching is shown in Table 5.2. Tracking results on Sequence 1~3 is shown in Figure 5.6~8.

Sequence	Tracked person	No. of Frames	Number of Lost Frames		Lost Percentage	
	ID No.	containing the target	Color-Based Kalman filter	Ours	Color-Based Kalman filter	Ours
No. 1	No. 1	199	15	0	7.54%	0
	No. 2	46	10	0	21.74%	0
No. 2	No. 1	199	19	0	9.55%	0
	No. 2	51	1	0	1.96%	0
	No. 3	30	3	0	10.00%	0
No. 3	No. 1	41	10	1	24.39%	0.244%
	No. 2	63	19	3	30.16%	4.76%
	No. 3	64	15	1	23.43%	1.56%
	No. 4	87	34	3	39.08%	3.45%

Table 5.2 Systematic Evaluation for Our method for Head Tracking Compared with Color-Based Tracking by Kalman filter

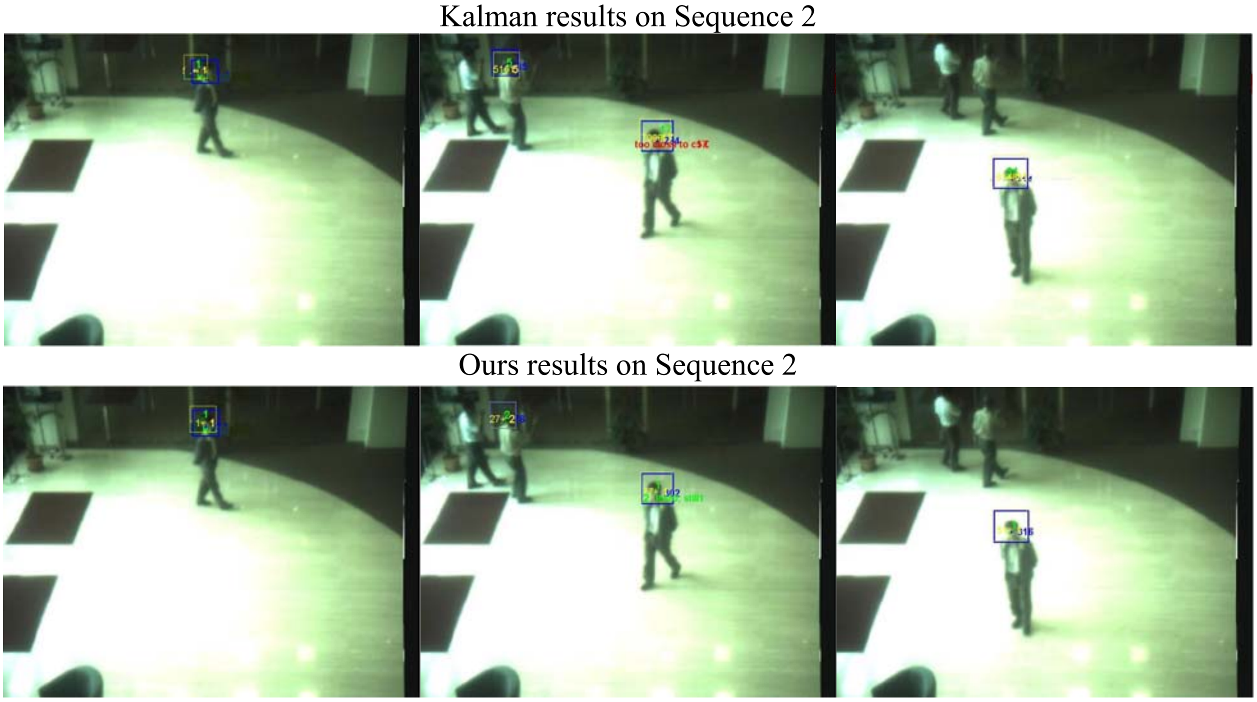


Kalman results on Sequence 1

Ours results on Sequence 1

Figure 5.6 Comparison between our method and a simple Kalman model using color-based matching for tracking on Sequence 1

Sequence 1 contains three persons: the person wearing a gray coat walks around in the upper area of the three pictures. The person with backpack and the one in pink enter the hall at the middle of this sequence. Most of Sequence 2 shows one person in gray coat walks around the hall. There were two persons walk into the scene and then walk out, which happens in a short period in Sequence 2. In Sequence 3, there is a person who walks straight from the left to the right of the scene, and then two girls walk side by side from the right to the left. Finally a person enters the scene from the left bottom and goes out of the gate in the end.

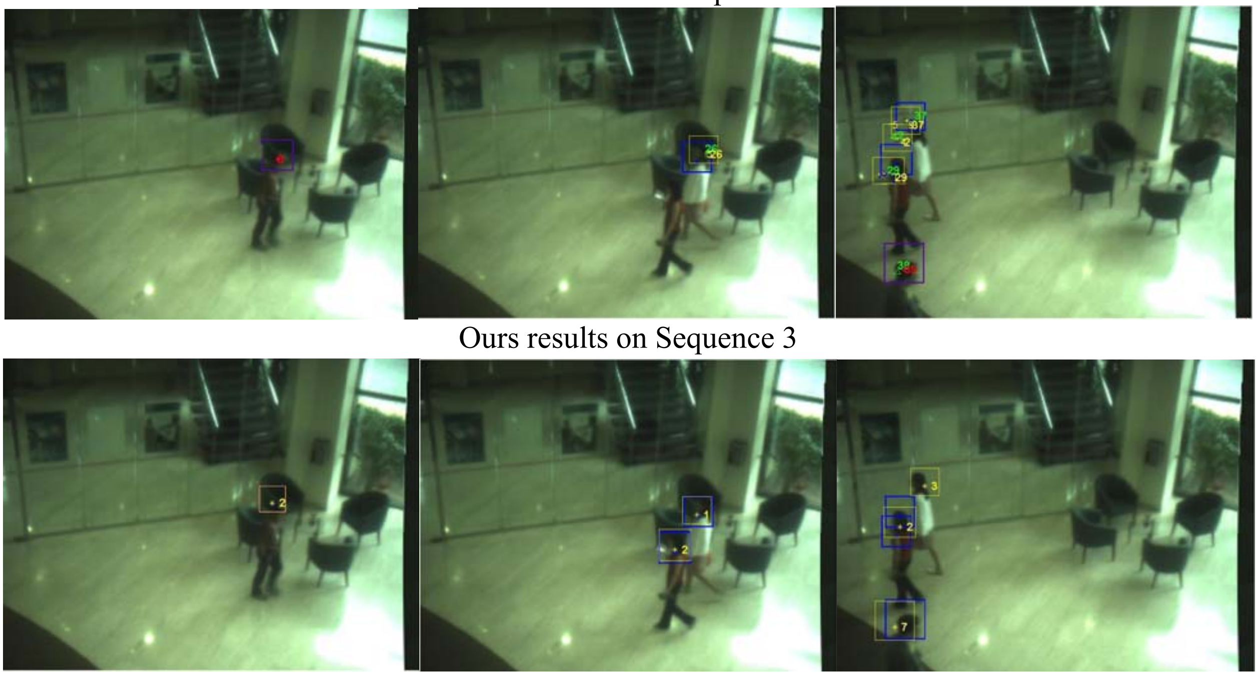


Kalman results on Sequence 2

Ours results on Sequence 2

Figure 5.7 Comparison between our method and a simple Kalman model using color-based matching for tracking on Sequence 2

In Figure 5.6~5.8, the red bounding boxes centered at red stars are human heads recognized by the system as newly appeared in the scene. (In this case, the red number indicates the ID assigned to the object by our system.) The yellow bounding boxes centered at yellow star shows object currently being tracked. (In this case, the yellow number indicates the ID of the object.) The stars are the human head centers organized by our system. It can be clearly seen that in the results of Color-based Kalman filter, the target is frequently lost and another new ID is assigned to it, so that the ID No. is very big afterwards, compared to that of the fusion tracking method. It shows that Kalman filter based on color fails in areas with the similar color as the human head and it also fails when the person changes his head pose, which leads to the change in the appearance of the head in the image. Also the region around the head changes a lot when the head is moving in clutter background. Another reason is that the low frame rate (about 8 frames per second) of our experiment sequences results in great distance from object position predicted by Kalman filter to the real object position. However, our method tracks the human heads well in the whole sequence.



Kalman results on Sequence 3

Ours results on Sequence 3

Figure 5.8 Comparison between our method and a simple Kalman model using color-based matching for tracking on Sequence 3

From the above results of our method for human tracking, we can see that our method works well even when a simple Kalman filter updated by color-based matching fails. It not only improves the performance of our stereo-based method for human head detection, but also obtains more accurate observation of the objects by integrating stereo, color and motion information in the consecutive frames. The results demonstrate that our tracking method for tracking successfully provides accurate motion information of each individual in the scene.

Chapter 6

Conclusion

In this paper, a stereo-based solution to the problem of detecting and tracking human individuals in crowds is proposed. The depth information allows us to estimate the size of human heads in the image. Based on such estimation, a process of scale-adaptive filtering is proposed to extract the evidence for the presence of human heads while suppress that of other objects. To suppress the spurious clues further, a virtual plane that is parallel to and above the ground surface with the average human height is established. The extracted clues whose 3D positions in the real space are far away to the virtual plane are suppressed. After that, a mean-shift algorithm is proposed to locate the bright blobs in the likelihood map for human head detection. Finally, the head of each individual in the crowd is tracked by fusing the information of motion color, and stereo adaptively. Promising results have been obtained from the experiments on real scene.

6.1 Contributions

Contributions of Stereo-based Human Detection: The main contributions of our stereo-based method for human head detection are: (1) propose the method of Object-Oriented Scale-Adaptive Filtering (OOSAF) to extract the evidence for head like objects from the stereo image, which is original; (2) propose a method of restoring the perspective of the view to suppress spurious clues which are much higher or lower than the average human height above the ground surface; (3) propose a mean-shift algorithm to detect and locate human heads in the likelihood map. The superiority of this method over other detection techniques on 2D images is that it can work well in crowded scenes where there is long-term and serious occlusion. Moreover, using the stereo information, the detection of human head becomes much more efficient. Another advantage of the proposed method is that it does not depend on background subtraction. So it is less sensitive to the shadows and other background changes. Besides, both stationary and moving objects can be detected by applying this method. To our knowledge, this research is the first attempt to the tough problem of human individual detection in crowded scene based on stereo and our method for head detection solves the problem of human individual detection in crowded scenes very well, which is proved by the experiment results. The proposed method has been tested on the real image sequences containing many crowd scenes. The statistics show that promising results have been achieved by the proposed method.

Contribution of the method for human tracking: The contribution of this method is providing a general framework to fuse color, motion, and stereo information for human tracking in crowds. As to our system, the implementation of the tracking not only obtains the temporal correspondences of the objects, but also raises the correct detections rate and suppresses the false detections rate. As to tracking methods, to locate the position of the target in the current frame accurately is crucial. More and more attention is given to fusion methods for data association, which proves to be more robust. However, not much research has been conducted using the stereo feature in sequence with many people in groups. Our fusion method for tracking guided by stereo information is original and proved to be more reliable and accurate compared to those use only one or two kinds of information. The results show this fusion method works better than those use less kinds of information.

6.2 Future Work: Group Behavior Understanding

The crowd information can be obtained with human head detection and tracking, including crowd density and major movements of human flows. Certain classification and estimation about human group can be made accordingly. The future work is discussed as follows:

Classification of Group Behavior: Clustering the human individuals into groups based on the similarity of motion and trajectory features. Then recognize the behaviors of the groups according to the learned patterns. The patterns of the group behaviors could be Queue, Human Flow, Talking Group, Riot and Suspicious Individuals in Crowds.

Crowd Density Estimation: The crowd density information is very important to crowd surveillance. When an area reaches an occupation level that is greater than the safety level, people's safety can be in danger. With the detected heads of individuals in the current frame, it is easy to measure the crowd density, and then to control the density of the crowd under a certain “safety level”.

Bibliography

[1] Beymer, D.; Konolige, K., “Real-Time Tracking of Multiple People Using Stereo”, *Proceedings of the IEEE Frame Rate Workshop*, Corfu, Greece, 1999

[2] Catlin, D., “Estimation, control and the discrete Kalman filter”, Springer Verlag inc, 1989

[3] Chen, Y.; Rui, Y.; Huang, T. S., “JPDAF Based HMM for Real-Time Contour Tracking, *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.1, pp. 543 - 550, 2001

[4] Comaniciu, D.; Meer, P., “Mean shift analysis and applications”, *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol.2, pp. 1197–1203, 1999

[5] Darrell, T.; Gordon, G.; Woodfill, J.; Harville, M., “Integrated person tracking using stereo, color, and pattern detection”, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 601 - 608, 1998

[6] Davies, A.C.; Yin, J.H.; Velastin, S.A., “Crowd monitoring using image processing,” *Electronics & Communication Engineering Journal*, pp. 37-47, Feb. 1995

[7] Eveland, C.; Konolige, K.; Bolles, R.C., "Background Modeling for Segmentation of Video-Rate Stereo Sequences", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 266 - 271, 1998

[8] Grzeszczuk, R.; Bradski, G.; Chu, M. H.; Bouguet, J.-Y., "Stereo Based Gesture Recognition Invariant to 3D Pose and Lighting", *Proceedings of IEEE Computer Vision and Pattern Recognition*, vol. 1, pp. 826-833, 2000

[9] Haritauglu, I.; Harwood, D.; Davis, L.S., "W^4S: A real-time system for detecting and tracking people in $2\frac{1}{2}$ D", *European Conference on Computer Vision,* pp. 222-227, 1998

[10] Haritaoglu, I.; Harwood, D.; Davis, L.S., "W^4: real-time surveillance of people and their activities", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, pp. 809-830, 2000

[11] Harvey, A.C, "Forecasting, structural time series models and the Kalman filter", Cambridge University Press, 1991

[12] Huang, Xiaoyu; Li, Liyuan; Sim, Terence, "Stereo-Based Human Head Detection From Crowd Scenes", *Proceedings of IEEE International Conference on Image Processing*, 2004

[13] Isard, M.; Blake, A., “Contour tracking by stochastic propagation of conditional density”, *Proceedings of European Conference on Computer Vision*, pp. 343-356, 1996

[14] Kalman, R.; Bucy, R., “New results in linear filtering and prediction theory”, *Journal of Basic Engineering, Transactions ASME Series D*, Vol. 83, pp 95-108, 1961

[15] Konolige, K., “Small Vision System: Hareware and Implementation”, *Proceedings of International Symposium on Robotics Research*, pp. 111-116, 1997

[16] Li, Liyuan; Ge, Shuzhi Sam; Sim, Terence; Koh, YingTing; Huang, Xiaoyu, “Object-Oriented Scale-Adaptive Filtering For Human Detection from Stereo Images”, *IEEE Conference on Cybernetics and Intelligent Systems*, 2004

[17] Lo, B.P.L.; Velastin, S.A., “Automatic congestion detection system for underground platforms”, *Proceedings of 2001 International Symposium on Video and Speech Processing*, pp. 158 -161, 2001

[18] Marana, A. N.; Costa, L. F.; Lotufo, R. A.; Velastin, S. A., “Estimating Crowd Density with Minkowski Fractal Dimension”, *IEEE International Conference On Acoustics, Speech, and Signal Processing*, 1999

[19] Marana, A. N.; Costa, L. F.; Lotufo, R. A.; Velastin, S. A., “On the Efficacy of Texture Analysis for Crowd Monitoring”, *International Symposium on Computer Graphics, Image Processing, and Vision*, Vol.6, pp. 3521 – 3524, 1998

[20] Marana, A.N.; Velastin, S.A.; Costa, L. F.; Lotufo, R.A., “Automatic Estimation of Crowd Density using Texture”, *Journal: Safety Science*, Vol. 28, pp. 165-175, 1998

[21] Morency, L L.-P.; Rahimi, A.; Checka, N.; Darrell, T., “Fast stereo-based head tracking for interactive environments”, *Proceedings of Conference on Automatic Face and Gesture Recognition*, pp. 375-380, 2002

[22] Nummiaro, K.; Koller-Meier, E.; Van Gool, L.J., “Object Tracking with an Adaptive Color-Based Particle Filter," *DAGM-Symposium Pattern Recog-nition*, pp. 353-360, 2002

[23] Okuma, K.; Taleghani, Ali; Freitas, N. de; Little, J. J.; Lowe, D. G., “A Boosted particle filter: Multitarget Detection and Tracking”, *The Eighth European Conference on Computer Vision*, Vol. 1, pp 28-39, 2004

[24] Russakoff, D.; Herman, M., “Head tracking using stereo,” *Proceedings of IEEE Workshop on Applications of Computer Vision*, pp. 254–260, 2000

[25] Velastin, S.A.; Yin, J.H.; Davies, A.C.; Vicencio-Silva, M.A.; Allsop, R.E.; Penn, A., “Automated Measurement of Crowd Density and Motion Using Image Processing”, *7th IEEE International Conference on Road Traffic Monitoring and Control*, pp. 127-132, 1994

[26] Vermaak, J.; Doucet, A.; Perez, P., “Maintaining Multimodality though Mixture Tracking”, *International Conference on Computer Vision*, Vol.2, pp1110 - 1116, 2003

[27] Zhao, L.; Thorpe, C. E., “Steteo- and Neural Network-Based Pedestrian Detection”, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 1, pp. 148-154, 2000

[28] Zhao, T.; Nevatia, R., “Bayesian human segmentation in crowded situations,” *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,* Vol. 2, pp. II - 459-66, 2003

[29] Zhao, T.; Nevatia, R., “Tracking Multiple Humans in Crowded Environment”, *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 406- 413, 2004

[30] Cupillard, F.; Avanzi, A.; Bremond, F.; Thonnat, M., “Video understanding for metro surveillance”, *IEEE International Conference on Networking, Sensing and Control*, Vol. 1, pp. 186-191, 2004