Principal Axis-Based Correspondence Between Multiple Cameras for People Tracking

Weiming Hu, Min Hu, Xue Zhou, Tieniu Tan (IEEE Fellow), Jianguang Lou
(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100080)
{wmhu, mhu, tnt, jglou}@nlpr.ia.ac.cn

Steve Maybank (IEEE Member)
(School of Computer Science and Information Systems, Birkbeck College, Malet Street, London WC1E 7HX)
sjmaybank@dcs.bbk.ac.uk

Abstract Visual surveillance using multiple cameras has attracted increasing interest in recent years. Correspondence between multiple cameras is one of the most important and basic problems which visual surveillance using multiple cameras brings. In this paper, we propose a simple and robust method, based on principal axes of people, to match people across multiple cameras. The correspondence likelihood reflecting the similarity of principal axis pairs is constructed according to the relationship between "ground-points" detected in each camera view and the intersections of principal axes detected in different camera views and transformed to the same view. Our method has the following desirable properties: (1) Camera calibration is not needed. (2) Accurate motion detection and segmentation are less critical due to the robustness of the principal axis-based feature to noise. (3) Based on the fused data derived from correspondence results, positions of people in each camera view can be accurately located even when the people are partially occluded in all views. The experimental results on several real video sequences from outdoor environments have demonstrated the effectiveness, efficiency and robustness of our method.

Keywords: Correspondence between multiple cameras, Principal axes, People tracking.

1. Introduction

In recent years, visual surveillance [33] using multiple cameras has attracted much attention in the computer vision community. This is because by using multiple cameras the area of surveillance is expanded and information from multiple views is extremely helpful to handle many issues such as occlusion etc. However, visual surveillance using multiple cameras also brings a number of problems such as camera installation [26], calibration of multiple cameras [13, 27], correspondence between multiple cameras [1-11], automated camera switching [9], and data fusion [28-31], etc. In this paper, we focus on correspondence between multiple cameras. Correspondence between multiple cameras involves at the same time instant finding correspondences between objects in the different image sequences. Only after correspondence between multiple cameras is well constructed, can the information from multiple cameras be fused. So, it is one of the most important and basic problems in visual surveillance using multiple cameras. As people activities are of key interest in monitored scenes, in this paper we mainly consider correspondence between multiple cameras for tracking people, and **improving the tracking of people when the people are within the common ground plane of the multiple views.**

Although correspondence of multiple cameras is a newly emergent research topic, in recent years some attempts have been made to investigate this problem. The existing methods for establishing correspondences can be classified [1, 2], according to the types of employed features, whether the cameras are calibrated or not, and whether the correspondences are region-based or point-based. The following sections describe existing methods in order to provide the context for our own work.

1.1. Region-based methods

Region-based methods generally regard people as regions and use the features of the regions to match people in multiple views. Color is a popular region cue to generate correspondence across views. Orwell *et al*. [3] and Krumm *et al*. [4] use color histograms to match people in different views. Mittal *et al*. [5] apply Gaussian color models to solve the problem of correspondence across multiple cameras.

Change *et al*. [1] establish cross camera correspondence by combining epipolar geometry, landmarks, height and color mapping between two cameras.

It is natural and simple to use color information to construct correspondence between multiple cameras. However, color-based correspondence across multiple cameras is highly unreliable. For one thing, it greatly relies on the colors of people's clothes. When different people have similarly colored clothes, this method may produce wrong correspondence. For another, viewpoint difference and lighting variations may cause the same person to be observed with different colors in different cameras. If a person's clothing is white at the front and black at the back, then the observations of the person in two views may be regarded as arising from two different people.

1.2. Point-based methods

A more applicable scheme for constructing multiple camera correspondence may be to match feature points of objects in each view based on geometric constraints. Feature point-based correspondence methods can be further divided into two sub-classes: 3D and 2D methods, according to the types of geometric constraints which are used.

(1) 3D methods

There are two strategies in 3D methods to establish correspondence:

- One is to transform all feature points into the same 3D space, and then match the feature points based on the principle that corresponding feature points in different views are projections of the same 3D point. In [6, 7], object centroids are taken as feature points and correspondence is established by estimating the corresponding 3D centroids in the world coordinate system. In [8], all cameras are calibrated and the 3D environment model is known beforehand. Correspondence across views is achieved for people who have similar estimated 3D locations.
- The second strategy uses 3D epipolar constraints for matching. In [9], only the relative calibration between neighboring cameras is used to derive epipolar constraints, and correspondence is established by matching a set of feature points along the midline of the upper part of a human body, based on the epipolar constraints.

The above strategies both need prior calibration of the cameras. Furthermore, feature points of a person extracted from different views do not always correspond to the same physical 3D point. This may make the correspondence of feature point pairs ambiguous.

(2) 2D methods

To overcome the disadvantages of 3D methods, some methods using 2D information have been presented to establish correspondence between multiple cameras. Khan *et al*. [2, 10] use the points located on feet to match people in multiple views, based on the homography constraint defined by the ground plane. However, in real applications, the points of people's feet may not be robustly or accurately detected or even are invisible due to occlusions.

Black *et al*. [11] provide the transfer error, based on a homography constraint, for the correspondence between the centroids in different views. The performance of this method deteriorates if only part of a person is visible or detected.

The current point-based methods aforementioned, either 3D or 2D, are easily influenced by noise. Accurate motion detection is required for these methods. If the motion is not well segmented or only part of a person is visible due to occlusion, the feature points may be unreliable and thus the performance of the correspondence is degraded.

1.3. Our method

In this paper, we propose a simple and robust method to match people in multiple views. In our method, the principal axis of each person, i.e. the symmetric axis of the human body, is detected in image planes and used to match people across views. Our method has the following main contributions:

- We use principal axes of people as features for the correspondence. Estimation of principal axes does not rely on accurate motion detection, because the influence of the error of motion detection is counteracted by the symmetrical distributions of the error along the principal axis.
- We propose a Least Median of Squares-based algorithm to detect principal axes of people under three situations: "isolated", "in a group", and "occluded".
- We find that the intersection of the principal axis of a person in a view and the line acquired by transforming the principal axis of this person from another view to the first view using a homography, corresponds the "ground-point" of this person in the first view. This "ground-point' is the intersection point of the principal axis and the ground plane in the first view.
- Based on the above property, we define the correspondence likelihood reflecting the correspondence similarity of principal axis pairs. Accordingly, the algorithm for matching between multiple cameras is presented.
- **Correspondence results are further used to improve the tracking results when the "ground points" of the tracked people are within the common ground plane of the multiple views. This fusion of data makes it is possible to robustly find and track the positions of people in different views, even when the people are partially occluded in all views.**

The remainder of this paper is organized as follows. Section 2 gives an overview of our method. Section 3 describes detection of principal axes and "ground-points" under different situations. Section 4 presents how principal axes are used to construct correspondence of people across multiple cameras. Section 5 covers how the information from multiple cameras is fused. Section 6 shows experimental results. The last section summarizes the paper.

2. Overview of Our Method

The motivation of this paper is to construct a robust method to match people across multiple cameras. We select principal axes of people as the feature for this matching. As foreground pixels corresponding to a person are symmetrically distributed along the principal axis, the errors of motion segmentation are distributed symmetrically. This reduces the influence of these errors on the detection of the principal axis of the person. So the principal axis feature is robust to noise.

As mentioned in Introduction, we find that the intersection of the principal axis of a person in a camera view and the line obtained by transforming, using the homography, the principal axis of the same person from another view to the first view is the "ground-point" of the person in the first view. So we can use the distance between the "ground-point" of a principal axis in a view and the intersection between this principal axis and the line obtained by transforming a principal axis from another view to the first view, to evaluate the degree of matching between these two principal axes in the two views.

Fig. 1 shows the process of detecting principal axes of people in a single camera. Based on the segmented foreground regions, people are distinguished from others, and are further classified into three categories: isolated people, a group of people, and people under occlusion. Principal axes are detected in these three situations. The tracking of people is based on their "ground-points'. In the detection of

"ground points" of people in a group or under occlusion, the results of prediction in the tracking module are employed.

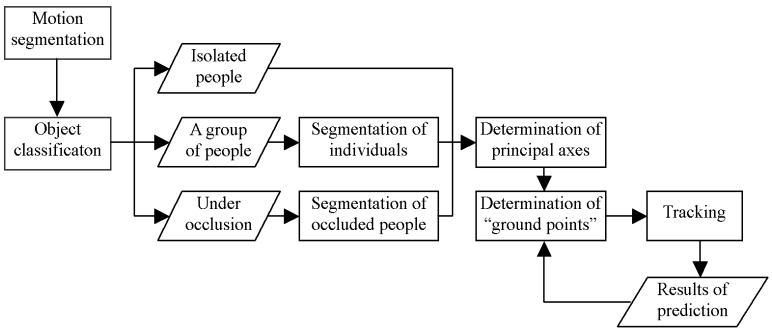


Fig. 1. Overview of principal axis detection in a single camera

Fig. 2 shows the process of people correspondence between multiple cameras. The correspondence between multiple cameras is based on the detection of principal axes of people and tracking of people in each single camera. In the matching of principal axes, a homography rather than camera calibration is used as the geometrical constraint. The correspondence depends on the intersection of a principal axis in a view and the line obtained by transforming a principal axis from another view to the first view. The correspondence results are fed back to the single camera tracking module, i.e. **the intersection of the principal axis of a person in one view and the line obtained by transforming the principal axis of the person from another view to the first view is used to update the "ground-point" of this person in the first view. Such feedback makes the tracking and correspondence robust (see Section 5 for more details).**

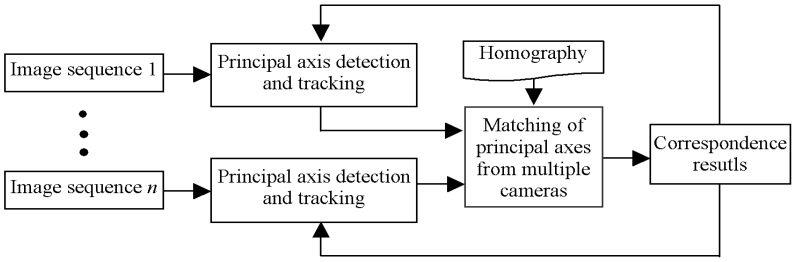


Fig. 2. Overview of people correspondence between multiple

For simplicity, we assume that people move in an upright pose and that the camera is oriented such that a vertical line in the scene projects to a line which is approximately vertical in the image. Thus, we only need to consider the case in which the principal axes are vertical in image planes. This assumption is reasonable in most surveillance applications and has been widely used [18, 19].

3. Detection of Principal Axes in a Single Camera

In this section, motion segmentation, object classification, detection of principal axes of people, tracking of people, and determination of "ground-points" in a single camera are described.

3.1. Motion segmentation and object classification

In this paper a simple background subtraction algorithm taken from [14] is used to extract foreground regions corresponding to moving objects. Other more adaptive and complex background subtraction methods [15-17] may be implemented to obtain better results. However, to verify the robustness of our method, we just use the simple algorithm.

Objects other than people, such as vehicles, may move in a monitored scene. It is necessary to use a simple classifier to distinguish people from other objects. We only consider the classification of people and vehicles. Based on the fact that a person's shape can be represented by its projection histogram [20], we propose a vertical projection histogram-based algorithm to distinguish people from vehicles. The vertical projection histogram is acquired by projecting foreground pixels onto the horizontal coordinate of the image. Let $I(x,y)$ be a binary image which represents a detected motion region, where x and y are respectively the horizontal and vertical coordinates. Let "*height*" and "*width*" be respectively the height and width of this motion region. The vertical projection histogram h is given by: $h(x) = \sum_{y=1}^{height} I(x,y)$, $x \in [1, width]$. The Y-coordinate of the vertical projection histogram is the total number of pixels with the

same horizontal coordinate.

We define the spread of a vertical projection histogram:

$$Spread = \frac{\sum_{x=1}^{width-1} |h(x+1) - h(x)|}{\sum_{x=1}^{width} h(x)}. \quad (1)$$

It is obvious that the vertical projection histogram of an isolated person is steeper than that of a vehicle, so the spread of an isolated person is higher than that of a vehicle. As the edge of the region of foreground pixels of a vehicle is smoother than that of a group of people, so the spread of a group of people is also higher than that of a vehicle.

If a foreground region is classified as people, we need to further decide whether it is an isolated person or a group of people by analyzing the number of significant peaks in its vertical projection histogram. This is detailed in Section 3.2.2.

3.2. Detection of principal axes

According to real applications, we consider detection of principal axes under three situations: "isolated", "in a group", and "under occlusion".

3.2.1. Principal axis of an isolated person

In this paper, we use the Least Median of Squares [21] to determine the principal axis of an isolated person, based on the global shape constraint that a human body is typically close to symmetrical around the principal axis. The principal axis is determined by minimizing the median of squared perpendicular distances between the foreground pixels and a vertical axis. Let $D(X_i,l)$ be the perpendicular distance between ith foreground pixel X_i and an axis l to be determined, as shown in Fig. 3. The principal axis L is estimated with: $L = \arg\min_l median_i \{D(X_i,l)^2\}$.

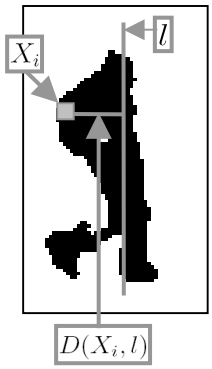


Fig. 3. Principal axis of an isolated person

3.2.2. Principal axes of people in a group

Two or more people, whose image motion regions overlap each other producing one foreground region, are regarded as a group. Determination of the principal axes of people in a group includes the following two stages:

- The whole region of a group is segmented into sub-regions, each of which corresponds an individual.
- The principal axis of each individual is determined.

The principal axis of a segmented individual can be determined as the same way as detecting the principal axis of an isolated person. So segmentation of individuals is critical in determination of principal axes of people in a group. Referring to [22], we use the vertical projection histogram which has been introduced in Section 3.1 to segment individuals, based on the fact a distinct peak region between two major valleys in the histogram corresponds to an individual. A distinct peak region should satisfy two conditions:

- All peaks within the distinct peak region

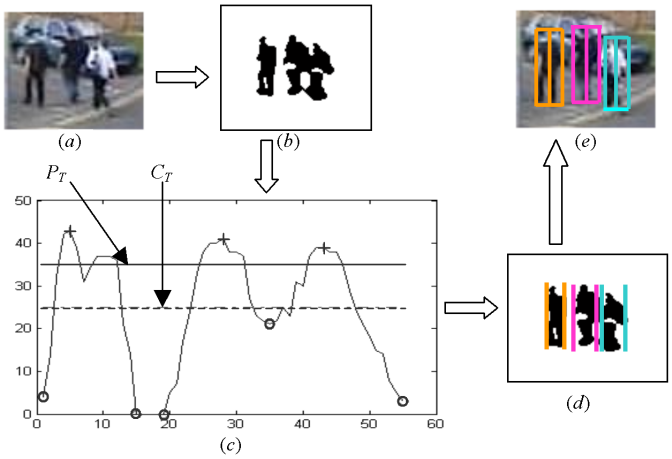


Fig. 4. Detection of principal axes of people in a group: (*a*) Input image. (*b*) Detected foreground region. (*c*) Vertical projection histogram. (*d*) Segmented individuals. (*e*) Principal axes.

are above a peak threshold (P_T)

- Two valleys of the distinct peak region must be lower than a valley threshold (C_T).

Thresholds P_T and C_T are selected empirically.

Fig. 4 shows an example of detection of principal axes of people in a group. The input image, the detected foreground region, and the corresponding vertical histogram are shown respectively in (*a*), (*b*), and (*c*). In (*c*), a sign "+" represents a peak, a sign "o" a valley, the solid line the peak threshold (P_T), and the dashed line the valley threshold (C_T). In this histogram, there are three distinct peak regions. According to these regions, three individuals are segmented as shown in (*d*) and their principal axes are correctly detected as shown in (*e*).

3.2.3. Principal axes of people under occlusion

To accurately detect the principal axis of an occluded person, it is necessary to segment the foreground pixels corresponding to the person from the whole foreground region. **The methods for color pixel classification for segmenting objects in general include appearance template-based ones [23], color histogram-based ones [34], and Kernel density-based ones [35]. The color template-based methods record the spatial color information of each pixel of each human body, and thus have high reliability in classifying foreground pixels, although there is a large amount of redundant information in the template. Considering the reliability, we select the color template-based method used in [23] to segment people under occlusion.** The color model (M_i) of object i consists of a color variable $C_i(X)$, which records the *rgb* color of each pixel X of object i, and an associated probability mask $P[M_i(X)]$, which records the likelihood of object i being observed at pixel X. A color model is initialized when a new object is tracked, and then updated in each new frame. The coordinates of pixel X are normalized to the current object position in image coordinates. The color of each pixel of the moving object is approximated with a Gaussian distribution. Let $I(X)$ be the observed color at pixel X. The probability density of $I(X)$ under the distribution in color model i ($p[I(X)|M_i(X)]$) is then acquired. During occlusion, segmentation of foreground pixels is formulated as a classification problem which is to determine the model to which each foreground pixel belongs. Using the Bayesian rule, the probability $P[M_i(X)|I(X)]$ of model i, given that $I(X)$ is observed, is determined. Pixel X is classified to the model m, if the probability of model m given pixel X is maximal.

After occluded people are segmented, their principal axes are determined as the same way as detecting principal axes of "isolated" persons.

3.2.4. Distinguishing between the three situations

In Sections 3.2.1, 3.2.2, and 3.2.3, we describe the methods for detecting principal axes of people under three situations: "isolated", "in a group", and "under occlusion". The remaining question is how to distinguish between these three situations. These situations can be distinguished by object correspondence relationships between consecutive frames in the tracking process. An object tracked in previous frames is defined as a "tracked object", and a motion region detected in the current frame is defined as a "detected object". The three situations are distinguished using the following principles:

(1) If only one "tracked object" which is a person in the previous frame corresponds to a "detected object", and only one significant peak region is detected in the vertical histogram of the "detected object", the "detected object" is classified as an isolated person.

(2) When more than one "tracked objects", each of which is a person, in the previous frame, correspond to a "detected object" in the current frame, the "detected object" is potentially a group of

people. In this case, we first use the method in Section 3.2.2 to segment the "detected object". If the method fails, we classify the case as "under occlusion" and use the method in Section 3.2.3 to detect the principal axes of the people under occlusion.

(3) If more than one "tracked objects" correspond to a "detected object" and these "tracked objects" include not only people but also objects other than people, we treat this case as "under occlusion".

3.3 Tracking

In Section 3.2.4, "tracked objects" and "detected objects" are defined. Tracking is in fact the construction of correspondence relationships between "tracked objects" in previous frames and "detected objects" in the current frame. A filter is used to predict the states of "tracked objects" in the current frame according to the states of the "tracked objects" in previous frames. Then, the predicted states are compared with the observations of the "detected objects", and the "detected object" corresponding to each "tracked object" is found. The states of the "tracked objects" in the current frame are updated using the corresponding "detected objects".

In this paper, the Kalman filter [32] is used to track people. The state of a person is (x,y,v_x,v_y) where (x,y) is the position of the person in the image plane and (v_x,v_y) is the velocity of the person. The observation of a person is its position (x,y). The position of an individual in a frame is evaluated with its "ground-point" on the image plane.

3.4. Detection of "ground-points"

In our method, the segmented foreground regions corresponding to individuals and the results of prediction in the tracking process are used to estimate the "ground-points" of detected principal axes.

For a detected principal axis, we find the intersection of the principal axis and the bottom line of the bounding box which contains the foreground pixels of the corresponding individual. If the distance between this intersection and the predicted position of this individual is small, this intersection is taken as the observation of the "ground-point" of the principal axis. Otherwise, the intersection of the principal axis and the vertical line from the predicted position to the principal axis is taken as the "ground-point".

We explain two points: (1) In most situations, the detected bounding box of an individual (especially an "isolated" person) is accurate, so the intersection of the bottom line of the bounding box and the principal axis can correspond well to the "ground-point". However, if the lower part of the individual is occluded or the color of the lower part of the individual is close to the background, the lower part of the individual may not be detected and this intersection cannot be regarded as the "ground-point". We use the distance between the predicted position and this intersection to check the occurrence of this situation. (2) The finally determined "ground-points" of people in the previous frame are very accurate as the correspondence results are used to update the estimated "ground-points" (see Section 5). This makes the predicted positions in the current frame are very close to the real "ground-points". So the estimated "ground-points" in the current frame are enough accurate even when the lower parts of people are occluded or cannot be segmented. (3) **In practice, the prediction is not critical since the change of positions of a move object between two consecutive frames is small. In fact, the "ground point" of a person in the previous frame is near to that in the current frame.**

4. Correspondence Between Multiple Cameras

In this section, we describe how detected principal axes are employed to establish the correspondence across multiple views given the homography constraint.

4.1. Homography recovery

To ensure the existence of the homography, it is necessary to assume that different views share a common dominant ground plane. This assumption is easily satisfied in most monitored scenes.

We define a 3×3 matrix: $H=\begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix}$. Let (x_i, y_i) and $(x_i^{'}, y_i^{'})$ be a pair of correspondence points on the ground plane in two views. They can be associated with *H*:

$$\begin{bmatrix} x_i^{'} \\ y_i^{'} \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}. \tag{2}$$

The homography *H* can be recovered from a set of static [12] or dynamic [13] correspondence points. In the paper, the ground plane homography is computed using several landmarks on the ground plane.

4.2. Geometrical relationship and correspondence likelihood

Before the correspondence likelihood function for evaluating the similarity of principal axis pairs across views is defined, the geometrical relationship of principal axis pairs across views is illustrated.

In Fig. 5, L_s^i is the principal axis of person *s* in view *i* and X_s^i is the "ground-point" of L_s^i. L_s is the principal axis of person *s* in 3D space, and X_s is the "ground-point" of L_s. g_s^i is the line acquired by projecting L_s onto the ground plane in 3D space from the direction of the view of camera *i*. Obviously, L_s^i is also the projection of g_s^i on image plane *i*. For person *k* in camera *j*, L_k^j, X_k^j, and g_k^j are similarly defined. Let H^{ij} be the ground plane homography from image plane *i* to image plane *j*. Let L_s^{ij} be the line in image plane *j*, obtained by transforming L_s^i from image plane *i* to image plane *j* using H^{ij}. Obviously, L_s^{ij} is also the projection of g_s^i on image plane *j*. Let Q_{sk}^{ij} be the intersection of L_s^{ij} and L_k^j. **It is obvious that if person *s* in camera *i* and person *k* in camera *j* correspond to the same person in 3D space, Q_{sk}^{ij} corresponds to the "ground-point" of the principal axis of this person in image plane *j*.**

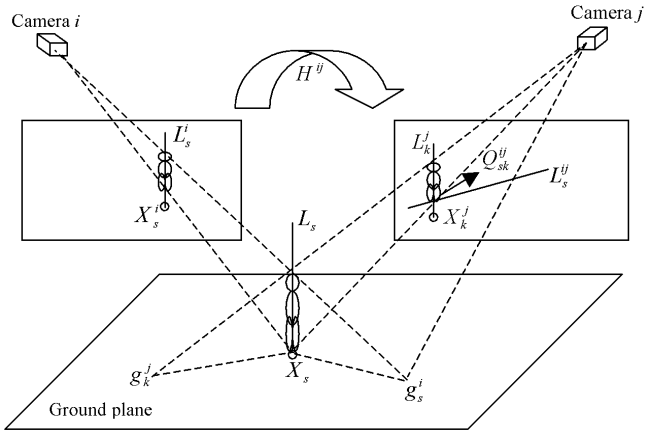


Fig. 5. Geometrical relationship between principal axes

Thus, the distance between the detected value of "ground-point" X_k^j and the intersection Q_{sk}^{ij} can be used to evaluate the correspondence likelihood for the pair of principal axes L_s^i and L_k^j. The less the distance is, the more likely the principal axes are matched.

In the same way, we can acquire the intersection Q_{ks}^{ji} in image plane *i*. The distance between the observation of "ground-point" X_s^i and the intersection Q_{ks}^{ji} also contributes to evaluate the correspondence likelihood between L_s^i and L_k^j. So, the function of correspondence likelihood between L_s^i and L_k^j is defined as:

$$\Re(L_s^i, L_k^j) = p(X_s^i \mid Q_{ks}^{ji}) p(X_k^j \mid Q_{sk}^{ij}). \tag{3}$$

To specify the likelihood values $p(X_s^i \mid Q_{ks}^{ji})$ and $p(X_k^j \mid Q_{sk}^{ij})$, without loss of generality, Gaussian distributions are assumed, as we think that the noise in the detection is approximated to the Gaussian distributions. Since Q_{ks}^{ji} and Q_{sk}^{ij} are close to the true values of "ground-points" in image planes, and X_s^i and X_x^j are observations of "ground-points" in image planes, we define $p(X_s^i \mid Q_{ks}^{ji})$ and $p(X_k^j \mid Q_{sk}^{ij})$ as:

$$p(X_s^i \mid Q_{ks}^{ji}) = 2\pi(\left|\Sigma_s^i\right|)^{-1/2} \exp\left\{-\tfrac{1}{2}(X_s^i - Q_{ks}^{ji})(\Sigma_s^i)^{-1}(X_s^i - Q_{ks}^{ji})^T\right\} \quad (4)$$

$$p(X_k^j \mid Q_{sk}^{ij}) = 2\pi(\left|\Sigma_k^j\right|)^{-1/2} \exp\left\{-\tfrac{1}{2}(X_k^j - Q_{sk}^{ij})(\Sigma_k^j)^{-1}(X_k^j - Q_{sk}^{ij})^T\right\} \quad (5)$$

where Σ_s^i and Σ_k^j are two covariance matrixes. Since coordinates *x* and *y* are independent, Σ_s^i is a diagonal matrix with two components $(\sigma_{xs}^i)^2$ and $(\sigma_{ys}^i)^2$, and Σ_k^j is diagonal with components $(\sigma_{xk}^j)^2$ and $(\sigma_{yk}^j)^2$.

The parameters of Σ_s^i and Σ_k^j are estimated with the mean of each distance between the observation of the "ground-point" and the corresponding intersection in each frame. In practice, Σ_s^i and Σ_k^j can be regarded as independent of image positions, i.e. $\Sigma_s^i = \Sigma^i$ and $\Sigma_k^j = \Sigma^j$.

To simplify the computation, we define the correspondence distance for principal axis pairs (D_{sk}^{ij}) according to (3), (4), and (5):

$$D_{sk}^{ij} = (X_s^i - Q_{ks}^{ji})(\Sigma^i)^{-1}(X_s^i - Q_{ks}^{ji})^T + (X_k^j - Q_{sk}^{ij})(\Sigma^j)^{-1}(X_k^j - Q_{sk}^{ij})^T. \quad \textbf{(6)}$$

The less D_{sk}^{ij}, the more likely the pair of principal axes (L_s^i, L_k^j) matches to each other.

4.3. Correspondence between multiple cameras

In this section, the defined correspondence distance is used to match people across multiple cameras. **In the following, we first present the matching algorithm for two views, and then explain how the algorithm generalizes to more than two cameras.**

It is assumed that, at time *t*, *M* people with principal axes L_1^i, L_2^i,…, L_M^i are observed from camera *i*, and *N* people with principal axes L_1^j, L_2^j,…, L_N^j from camera *j*. Our correspondence algorithm is to find the pairs of axes where the sum of correspondence distance values of the pairs is minimum. In this way, the principal axes are matched as a whole to avoid the potential errors caused by simply choosing the pair with the minimum distance.

The major steps of the correspondence algorithm are listed as follows:

Step 1: People principal axes detected in two different views are combined pairwise. Without loss of generality, it is assumed that $M \le N$. All principal axes in View *i* pair any of *M* principal axes in View *j*. Then there are $M \times N$ combination modes. For each combination mode *k*, it is assumed that principal axes L_1^i, L_2^i,…, L_M^i pair respectively off principal axes $L_{k_1}^j$, $L_{k_2}^j$,…, $L_{k_M}^j$, and then pair set $\theta_k = \{(L_1^i, L_{k_1}^j), (L_2^i, L_{k_2}^j)..., (L_{k_M}^i, L_{k_M}^j\}$ is found.

Step 2: For each pair $\{m,n\}$ in pair set θ_k, its correspondence distance D_{mn}^{ij} is computed. Then, it is checked that whether pair $\{m,n\}$ satisfies the constraint $D_{mn}^{ij} < D_T$, where D_T is a predefined threshold decided empirically to classify true and false correspondence pairs. If not so, pair $\{m,n\}$ is deleted from pair set θ_k. Then, pair set θ_k contains pairs satisfying the constraint.

Step 3: All pair sets with maximum number (*l*) of pairs are selected, and represented with

$\{\Theta = \{\Theta_k = (L^i_{k_1}, L^j_{k_1}), (L^i_{k_2}, L^j_{k_2}), ..., (L^i_{k_l}, L^j_{k_l})\}\}$.

Step 4: for each pair set in pair sets Θ, we look for the pair set (λ) with the minimum sum of correspondence distance values: $\lambda = \arg\min_k(\sum_{w=1}^{l}(D^{(i,j)}_{(k_w,k_w)}))$. All axis pairs in pair set Θ_λ are the matched ones.

Step 5: The pairs in pair set Θ are labeled.

For more than two cameras, cameras can be combined pairwise. Each pair of cameras, which have a common ground plane area, is corresponded using the above two camera correspondence algorithm. If there is inconsistency of correspondence between the camera pairs, the correspondence with less correspondence distance defined by (6) is selected.

5. Fusion of Data from Multiple Cameras

After all matched pairs are found, we can use the correspondence information to improve the tracking results in each single camera view **when tracked people are within the common ground plane of the multiple views**. As the principal axis of a person in each view can be detected robustly and accurately, for two cameras, the intersection of the principal axis of the person in one view and the line obtained by transforming the principal axis of the person from another view to the first view is robust and accurate to correspond to the real "ground-point" of this person in the first view. We use this intersection to update the former observation of the "ground-point" detected in the first camera. **As shown in Fig. 5, on the condition that principal axes L^i_s and L^j_k correspond to the same person, the intersection Q^{ij}_{sk} is used to more accurately estimate the "ground point" of L^j_k. So even when the "ground point" of the person is invisible (occluded or not detected) in both of the views, the intersection can be found, and thus the "ground point" of the person and accurately located in both of the views.**

For more than two cameras, there may be two or more such intersections in the first view for the person. If so, the mean of these intersections is selected as the "ground point" of the person.

The accuracy of the detected "ground-point" in the current frame insures that, if the person is occluded in the next frame, its "ground-point", which is determined using the predicted position and the detected principal axis, is accurate enough to make the correspondence correct.

Therefore, due to detection of principal axes, prediction, correspondence between multiple cameras and date fusion of multiple cameras, positions of people in different views can be robustly located and tracked, even if the people are "in a group" or "under occlusion".

6. Experiments

To verify our method, we have performed a number of experiments on our own NLPR database and the open PETS2001 database[1]. Furthermore, some comparisons have been implemented.

In our experiments, a tracked person is represented with a colored bounding box. Different persons are labeled with different colors. The vertical line within the bounding box of the person is the principal axis of the person, and the intersection of the principal axis and the bottom line of the bounding box is **the "ground-point" of the person, which is updated by the corresponding "intersection", as introduced in Section 5. Namely, the bottom edge line of the bounding box is determined by the "ground point" of the person, while the top, left, and right edge lines of the bounding box are determined by the foreground pixels of the person.**

[1] www.visualsurveillance.org/PETS2001

In our experiments, the spread threshold is selected to be 0.1. The valley threshold value (C_T) is selected as the mean value of the entire histogram. The peak threshold value (P_T) is selected as 80% of the height of the foreground pixel region. Threshold D_T for correspondence distance is empirically set to 5.

6.1. Results on NLPR database

Our NLPR video sequences are captured from outdoor environments, including Dataset 1 with two fixed cameras, and Dataset 2 with three fixed cameras.

In Dataset 1, each video sequence consists of 8,000 frames. For this database, people are correctly detected, matched and tracked in two views in most cases, except the following two cases:

- People in the image plane are too small to be detected;
- Very serious occlusion exists and models of people cannot be acquired beforehand due to few frames in which the people have appeared.

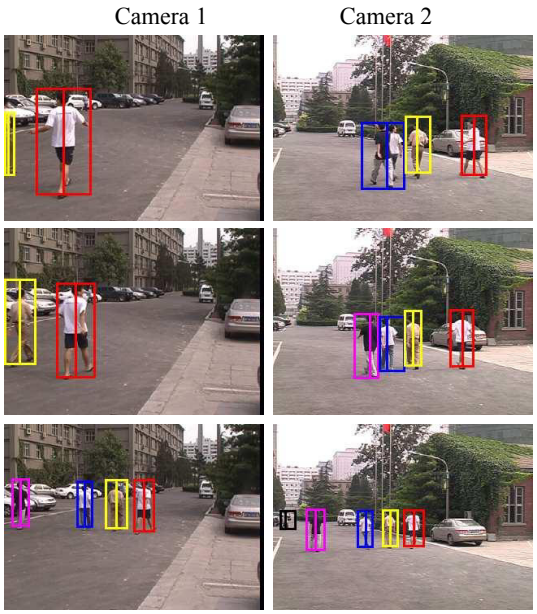


Fig. 6. Tracking and correspondence of multiple people with two cameras: From the top to the bottom, the frame numbers are respectively 3286, 3297, and 3380.

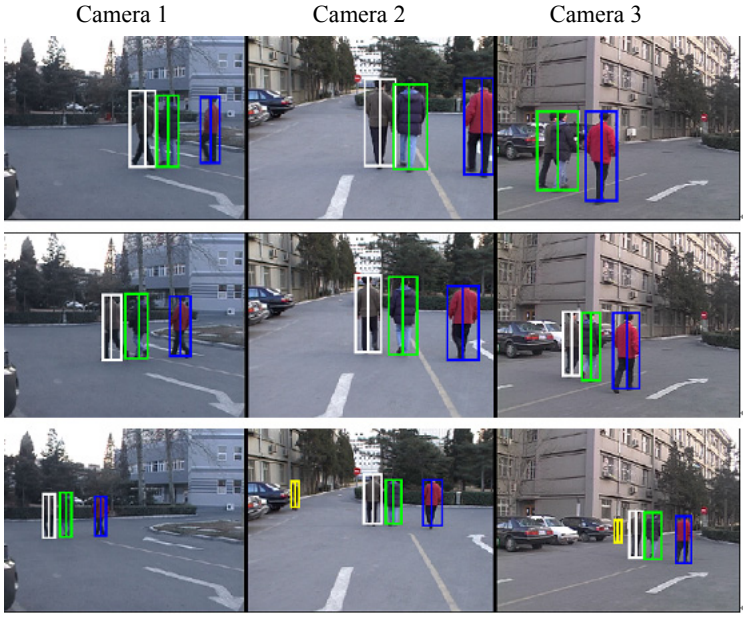


Fig. 7. Tracking and correspondence of multiple people with three cameras: From the top to the bottom, the frame numbers are respectively 33, 54, and 133

Fig. 6 illustrates some frames of tracking and correspondence results from Frame 3280 to Frame 3660. In this portion of video sequences, four people enter the field of the two camera views in succession and then leave. At Frame 3286, two people enter the second camera view so closely together that the method for detecting principal axes of people in a group fails to segment them apart. Furthermore, the duration that they have appeared in the view is too short to construct their models, and thus the method for detecting principal axes of occluded people is unusable. So they are tracked as an individual. However, they are successfully segmented and tracked after some frames when they are not so close but still in a group (see Frame 3297). From this figure, we can see that the principal axes of these people are successfully detected, people are matched correctly between the two cameras, and their positions in the image planes are exactly located. **(*Please see the supplemental video Example1.avi.*)**

For Dataset 2, people are also correctly tracked and corresponded in all the three views. Fig. 7 illustrates some frames of results from Frame 1 to 160. In this portion of sequences, four people enter the fields of the three camera views in succession and then leave. As shown in Frame 33, two people enter the third view so closely together that they cannot be segmented apart, so firstly they are tracked as an individual in the third view. However, they are successfully segmented and

tracked after some frames (see Frame 54). (*Please see the supplemental video Example2.avi.*)

6.2. Results on PETS2001 database

The PETS2001 database is the only open database available currently for the research of visual surveillance. Many algorithms [10, 11, 24, 25] have been evaluated on the database. We select dataset1 in the database for testing. Dataset1 consists of two video sequences captured by two static cameras in outdoor environments. In the whole sequence, all people having appeared are correctly matched and tracked in the two camera views using our method. Below shows and explains results of tracking and correspondence under the situations of "in a group" and "under occlusion". To illustrate the results of tracking clearly, only relevant parts of each image are displayed in the following figures.

Fig. 8 illustrates some frames when a moving person and a moving vehicle occlude each other from Frame 555 to Frame 601. In this sequence, a person and a vehicle move towards each other, meet, and then separate. At Frame 560, the lower part of the person is occluded by the vehicle in View 1. After some time, occlusion also happens in View 2, but now the person occludes the vehicle. During these occlusions, features of centroids and colors are no longer reliable for tracking and matching. In contrast, this person is still well tracked and matched using our principal axis-based method. Furthermore, the location of this person in each view is still accurately estimated, even when the "ground-point" of the person is occluded by the vehicle for example in Frames 560 and 578. **(*Please see the supplemental video Example3.avi.*)**

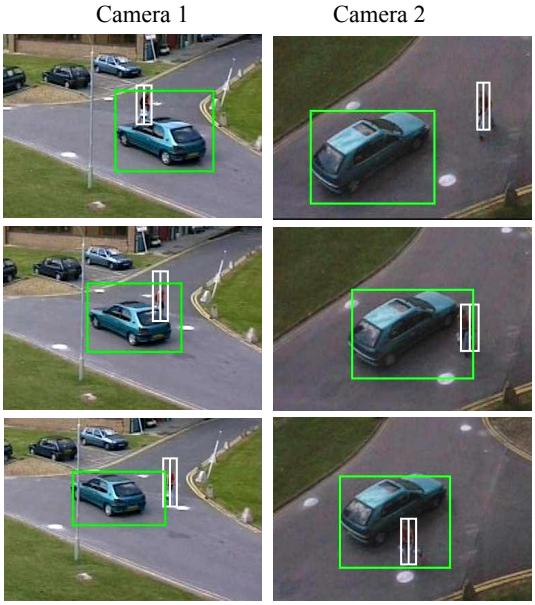


Fig. 8. Tracking and correspondence of a person occluded by a vehicle with two cameras: From the top to the bottom, the frame numbers are respectively 560, 578, and 590.

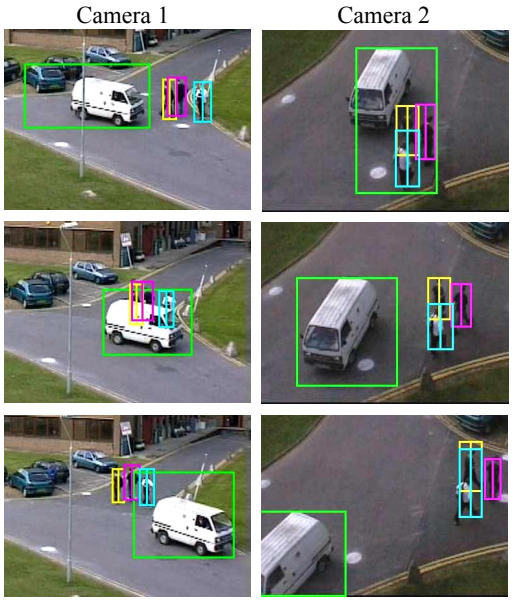


Fig. 9. Tracking and correspondence of a group of people through occlusion: From the top to the bottom, the frame numbers are respectively 878, 908, and 940.

Fig. 9 shows an example of tracking and matching a group of people even through occlusion. In this example, a group of three people and a vehicle move towards, meet and occlude each other, and then continue forward. It is noted that two persons in the group are both dressed in similar black. This makes it difficult to decide which person in one view corresponds to which person in another view using only color information, e.g. color histogram [3]. Furthermore, people interact with each other. For example, **in many frames, the base of one of them is occluded by another person along the vertical direction in View 2, and in View 1 the bases of all of them are simultaneously occluded by a vehicle.** This introduces extra difficulty for the processes of segmentation, tracking, and matching. In this case,

traditional feature points are very difficult to extract and not reliable for matching. However, using our principal axis-based method, **individuals in the group are correctly segmented**, tracked, and matched. **(*Please see the supplemental video Example4.avi.*)**

6.3. Comparison

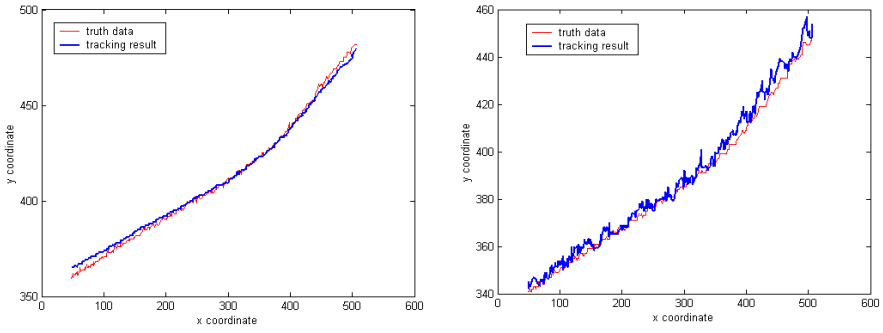
As discussed in Introduction, the existing methods for correspondence between multiple cameras can be classified into region-based ones and point-based ones. Color is the typical feature in region-based correspondence. As color is highly unreliable for correspondence, it is seldom used alone, but associated with other features. So we only compare our method with point-based correspondence method.

It is difficult to directly compare point-based methods with our method with respect to correspondence results, as different methods have different geometric constraints and applications. Although correspondence algorithms differ, the output of most tracking methods is motion trajectories of objects. Thus, we take the object trajectories as the basis of a comparison between our method and point-based tracking. We use centroids (**i. e. centers of the foreground regions corresponding to individuals or isolated persons**) as the chosen points because centroids are the most popular feature points. **The algorithm that we are comparing with is derived from [11], since the geometrical constraint in [11] is similar to ours. We compare the trajectories in the image plane.** Since there is no principle to directly determine which trajectory is more accurate, **we compare the centroid trajectory with the truth centroid trajectory obtained manually, and compare the "ground point" trajectory with the truth "ground point" trajectory obtained manually too. We measure the error of the centroid trajectory with the mean of the distance between the estimated centroid and the true centroid in each frame, and the error of the "ground point" trajectory with that between the estimated "ground point" and the true "ground point".**

(*a*) (*b*)

Fig. 10. Trajectories for comparison (the red ones are acquired using our method, and the blue ones are centroid trajectories): (*a*) Trajectories in View 1. (*b*) Trajectories in View 2.

A comparison is implemented on the sequence of the PETS2001 dataset 1 from Frame 2127 to Frame 2150. In this sequence a person walks along a road in two views. The trajectories in the two views are shown in Fig. 10. We select the comparison results in Camera 1, shown in Fig.11, for illustration. The results of comparison in Camera 2 are similar to those in Camera 1. The results shown in (*a*) are the comparison between the trajectory obtained using our method and the true "ground-point" data. The trajectory error for our method is 3.2 pixels. The results shown in (*b*) are the comparison between the trajectory acquired by tracking centroids and the true centroid data. The trajectory error for the centroid trajectory is 5.8 pixels. From Fig.11, we can see that the trajectory acquired using our method is more accurate than the



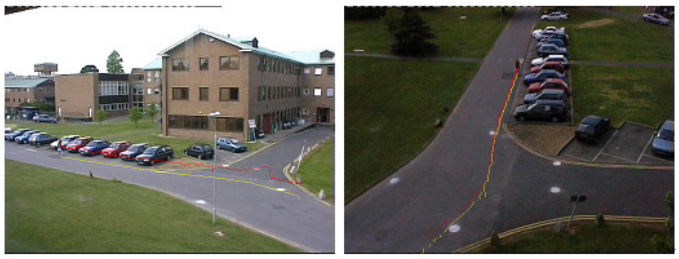
(*a*) E=3.2 (*b*) E=5.8

Fig. 11. Comparison: (*a*) Trajectory acquired using our method and true data. (*b*) Centroid trajectory and true data.

centroid trajectory. In addition, the trajectory acquired using our method is much smoother than the centroid trajectory.

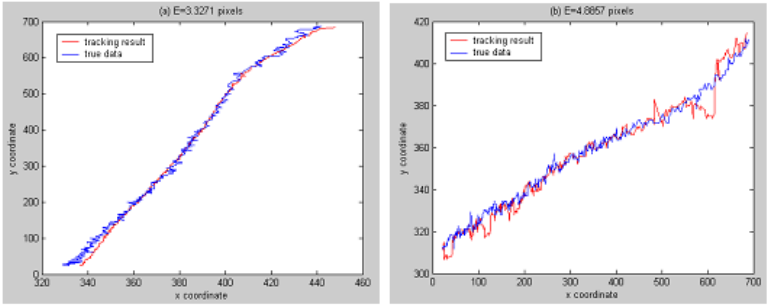
A Similar comparison is implemented on the sequence of the PETS2001 dataset 1 from Frame 250 to Frame 625. This sequence covers the sequence shown in Fig. 8. There exists occlusion between a person and a vehicle. The acquired trajectories in the two views are shown in Fig. 12. Fig. 13 shows the comparison result. As shown in (*a*), the trajectory error for our method is 3.3 pixels. As shown in (*b*), the trajectory error for the centroid trajectory is 4.9 pixels. From Fig 13, we can see that, when the person is occluded, estimated centroids are far apart from the true centoids, however the estimated "ground points" are near to the true "ground points.

Both of the above examples show that our principal axis-based method is more robust and efficient than the centroid-based one.



(a) (b)

Fig.12. Trajectory for comparison (the yellow ones are acquired using our method, and the red ones are centroid trajectories): (a) Trajectories in View 1. (b) Trajectories in View 2



(a) E=3.3 (b) E=4.9

Fig.13. Comparison: (a) Trajectory acquired using our method and true data. (b) Centroid trajectory and true data

7. Conclusions

In this paper, we have proposed a novel principal axis-based method for matching people across multiple cameras. In our method, camera calibration is not needed and there is less sensitivity to errors in motion detection. Principal axes and "ground-points" of people in each single camera view can be well detected even when the people are "in a group" or "under occlusion". Our algorithm for correspondence between multiple cameras is based on the fact that the intersection of the principal axis of a person in a view and the transformed principal axis of this person from another view corresponds to the "ground-point" of this person in the first view. This intersection is used to update this person's "ground-point" detected in the single view. This makes it possible to locate people accurately in the image plane even when they are partially occluded in all views. Our method has been tested on several real video sequences from the NLPR database and the PETS 2001 database. The experimental results have demonstrated the effectiveness, efficiency and robustness of our method.

In the future work, we will consider non-planar ground surfaces and make use of appearance information to pair up nearby people which could be confused.

References

1. T. H. Chang, S. Gong, and E. J. Ong, "Tracking Multiple People under Occlusion Using Multiple Cameras", *in Proc. of British Machine Vision Conference*, Bristol UK, September 2000, pp. 566-575.
2. S. Khan and M. Shah, "Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapping Fields of View", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 10, October 2003, pp.1355-1360.
3. J. Orwell, P. Remagnino, and G. A. Jones, "Multiple Camera Color Tracking", *in Proc. of IEEE International Workshop on Visual Surveillance*, Fort Collins, Colorado, June 1999, pp. 14-24.
4. J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-Camera Multi-Person Tracking for EasyLiving", *in Proc. of IEEE International Workshop on Visual Surveillance*, Dublin Ireland, July 2000, pp. 3-10.
5. A. Mittal and L. S. Davis, "M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene Using Region-Based Stereo", *in Proc. of European Conference on Computer Vision*, Copenhagen Denmark, May 2002, pp. 18-36.

6. H. Tsutsui, J. Miura, and Y. Shirai, “Optical Flow-Based Person Tracking by Multiple Cameras”, *in Proc. of IEEE Conference on Multisensor Fusion and Integration in Intelligent Systems*, Baden-Baden, August 2001, pp. 91-96.
7. A. Utsumi, H. Mori, J. Ohya, and M. Yachida, “Multiple Human Tracking Using Multiple Cameras”, *in Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, Kaiser, April 1998, pp. 498-503.
8. P. Kelly, A. Katkee, D. Kuramura, S. Moezzi, S. Chatterjee, and R. Jain, “An Architecture for Multiple Perspective Interactive Video”, *in Proc. of ACM Multimedia*, CA USA, November 1995, pp.201-212.
9. Q. Cai and J. K. Aggarwal, “Tracking Human Motion in Structured Environments Using a Distributed-Camera System”, *IEEE Trans. on Pattern Recognition and Machine Intelligence*, November 1999, Vol. 21, No. 11, pp. 1241-1247.
10. S. Khan, O. Javed, and M. Shah, “Tracking in Uncalibrated Cameras with Overlapping Field of View”, *in Proc. of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Hawaii USA, December 2001, pp. 84-91.
11. J. Black and T. Ellis, “Multi-Camera Image Tracking”, *in Proc. of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Hawaii USA, December 2001, pp. 68-75.
12. K. J. Bradshaw, L. D. Reid, and D. W. Murray, “The Active Recovery of 3D Motion Trajectories and Their Use in Prediction”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 3, March 1997, pp. 219-234.
13. L. Lee, R. Romano, and G. Stein, “Monitoring Activities from Multiple Video Streams: Establishing a Common Coordinate Frame”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, August 2000, pp. 758-767.
14. C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “Pfinder: Real-Time Tracking of the Human Body”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, July 1997, pp. 780-785.
15. W. E. L. Grimson, C. Stauffer, L. Lee, and R. Romano, “Using Adaptive Tracking to Classify and Monitor Activities in a Site”, *in Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, CA USA, June 1998, pp. 22-31.
16. C. Stauffer and W. E. L. Grimson, “Learning Patterns of Activity Using Real-Time Tracking”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, August 2000, pp.747-757.
17. A. Elgammal, D. Harwood, and L. Davis, “Non-Parametric Model for Background Subtraction”, *in Proc. of European. Conference on Computer Vision*, Dublin Ireland, June 2000, pp. 751-767.
18. T. Zhao, R. Nevatia, and F.Lv, “Segmentation and Tracking of Multiple Humans Complex Situations”, *in Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Kauai Hawaii, December 2001, pp.194-201.
19. A. M. Elgammal and L. S. Davis, “Probabilistic Framework for Segmenting People under Occlusion”, *in Proc. of IEEE International Conference on Computer Vision*, Vancouver Canada, July 2001, pp. 145-152.
20. S. Stillman, R. Tanawongsuwan, and I. Essa, “A System for Tracking and Recognizing Multiple People with Multiple Cameras”, *in Proc. of International Conference on Audio and Video-Based Biometric Person Authentication*, Washington DC, March 1999, pp. 96-101.
21. Y. Yang and M. Levine, “The Background Primal Sketch: an Approach for Tracking Moving Objects”, *Machine Vision and Applications*, Vol. 5, 1992, pp.17-34.
22. I. Haritaoglu, D. Harwood, and L.S. Davis, “W4: Real-Time Surveillance of People and Their Activities”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, August 2000, pp. 809-830.
23. A. Senior, “Tracking People with Probabilistic Appearance Models”, *in Proc. of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Copenhagen Denmark, June 2002, pp. 48-55.
24. Q. Zhou and J. K. Aggarwal, “Tracking and Classifying Moving Objects from Video”, *in Proc. of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Hawaii USA, December 2001, pp. 52-59.
25. L. M. Fuentes and S. A. Velastin, “People Tracking in Surveillance Applications”, *in Proc of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Hawaii USA, December 2001, pp. 20-27.
26. I. Pavlidis, V. Morellas, P. Tsiamyrtzis, and S. Harp, “Urban Surveillance System: from the Laboratory to the Commercial World”, *Proceedings of the IEEE*, Vol. 89, No. 10, 2001, pp. 1478-1497.
27. G. P. Stein, "Tracking from Multiple View Points: Self-Calibration of Space and Time", *in Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 1999, Vol. I, pp. 521-527.
28. S. L. Dockstader and A. M. Tekalp, “Multiple Camera Tracking of Interacting and Occluded Human Motion”, *Proceedings of the IEEE*, Vol. 89, No. 10, 2001, pp. 1441-1455.
29. R. T. Collins, A. J. Lipton, H. Fujiyoshi, and T. Kanade, “Algorithms for Cooperative Multi-Sensor Surveillance”, *Proceedings of the IEEE*, Vol. 89, No. 10, 2001, pp. 1456-1477.
30. V. Kettnaker and R. Zabih, “Bayesian Multi-Camera Surveillance”, *in Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 1999, pp. 253-259.
31. A. Utsumi, H. Mori, J. Ohya, and M. Yachida, “Multiple-View-Based Tracking of Multiple Humans”, *in Proc. of International Conference of Pattern Recognition*, 1998, pp. 197-601.
32. G. Welch and G. Bishop, “An Introduction to the Kalman Filter”, from http://www.cs.unc.edu, UNC-ChapelHill, TR95-041, November 2000.
33. W. M. Hu, T. N. Tan, L. Wang, and S. J. Maybank, “A Survey on Visual Surveillance of Object Motion and Behaviors”, *IEEE Trans. on Systems, Man and Cybernetics, Part C: Applications and Reviews*, Vol. 34, No. 3, 2004, pp. 334-352.
34. S. J. McKenna, S. Jabri, Z. Duric, and A. Rosenfeld, “Tracking Groups of People”. *Computer Vision and Image Understanding*, Vol. 80, No. 1, pp. 42–56, October 2000.
35. A. M. Elgammal and L. S. Davis, “Probabilistic Framework for Segmenting People under Occlusion”, *In Proc. of IEEE International Conference on Computer Vision*, pp. 145-152, July 2001.