Tracking Multiple People with a Multi-Camera System

Ting-Hsun Chang, Shaogang Gong
Department of Computer Science, Queen Mary, University of London
London, E1 4NS, UK {cth, sgg}@dcs.qmw.ac.uk

Abstract

We present a multi-camera system based on Bayesian modality fusion to track multiple people in an indoor environment. Bayesian networks are used to combine multiple modalities for matching subjects between consecutive image frames and between multiple camera views. Unlike other occlusion reasoning methods, we use multiple cameras in order to obtain continuous visual information of people in either or both cameras so that they can be tracked through interactions. Results demonstrate that the system can maintain people's identities by using multiple cameras cooperatively.

1. Introduction

Tracking moving people in an indoor environment is of interest in a number of applications such as visual surveillance, human-computer interface and video conferencing. Occlusion is a significant problem which can not be ignored because identities of people can become ambiguous. An example of an occlusion scenario is shown in Figure 1. This paper attempts to solve the occlusion problem in human tracking by using multiple uncalibrated static and widely-separated cameras.

Figure 1. People viewed from two widely separated cameras. The task is to track them with identities even when occlusion is present.

Different solutions to the occlusion problem in human tracking have been proposed. Rosales and Sclaroff [16] used Kalman filters and Khan and Shah [11] used colour. However, neither of these methods work for all cases. Recently, Haritaoglu et al. [6] implemented a real-time human-tracking system W^4 and suggested using a multi-camera system to analyse the occlusions.

Using multiple cameras to solve the occlusion problem, the cameras are separated widely in order to obtain visual information from wide viewing angles and offer a possible 3D solution [12]. The system needs to pass the subjects identities across cameras when the identities are lost in a certain view by matching subjects across camera views. Therefore, the system needs to match subjects in consecutive frames of a single camera and also match subjects across cameras in order to maintain subject identities in as many cameras as possible. Although this cross view correspondence is related to wide baseline stereo matching, traditional correlation based methods fail due to the large difference in viewpoint [14]. Because the variation is large between two camera views, the features used for matching should be view-independent or transformed to a suitable value for different cameras. To this end, Collins et al. [3] use the trajectory and normalised colour histogram of an object. Chang et al. [2] estimate the subjects' apparent height and apparent colour across cameras. This matching can also be done by employing the geometry of multiple views such as epipolar geometry [1] and homography [12] or scene knowledge such as landmarks [2]. However, these feature-based matching methods can be unreliable due to the ambiguous positions of the extracted features resulting in inconsistencies over time or conflict with each other. A framework is required to combine multiple visual modalities, or cues, to make the matching more reliable. Bayesian Networks [13, 10] provide such a framework which enable the full set of possible matching assignments to be simultaneously considered in a consistent and probabilistic manner. This Bayesian modality fusion method is related to the work of Toyama and Horvitz [17]. We also apply this fusion method to match subjects between consecutive frames from a single camera. Note that the method we present in this paper also can be used to track and follow multiple people as they move through the Field Of Views (FOVs) of different cameras.

In order to track individuals continuously, the system assigns an identity to a new detected subject and keeps tracking it with this identity. If this subject has already appeared in the other cameras or loses the identity during tracking,

the system then passes identity and re-assigns it to this subject by matching subjects across camera views. Thus, the tracking has two different modes: Single Camera Tracking (SCT) matching subjects between consecutive frames and Multiple Camera Cooperative Tracking (MCCT) matching subjects across cameras.

2. Bayesian Networks for Building Correspondence

In this section, we first define our problem. Then we explain the use of Bayesian networks to fuse multiple modalities to solve the correspondence problem. Firstly, we constrain the maximum number of subjects in each image to be m. To match subjects between 2 images, I_i and I_j, instead of matching each single subject independently which might result in conflicting results, we evaluate the matching globally, i.e. consider the matching for all subjects simultaneously. In each combination of assignments, every subject in I_i is assigned a corresponding subject in I_j. After applying the uniqueness constraint, i.e. a subject in I_j is allowed to be assigned to one and only one subject in I_i, there could be $m!$ possible assignment combinations, $A_\alpha = \{A_1, \cdots, A_{m!}\}$. Given the visual evidence **e** from all cameras which might be uncertain and incomplete, our goal is to find a most appropriate assignment combination which maximises the posterior:

$$\max_{\alpha \in \{1,\cdots,m!\}} p(A_\alpha|\mathbf{e}) \quad (1)$$

We employ Bayesian networks to probabilistically infer the correspondence of people in two images. The networks can capture the dependencies between the correspondence of the subjects between two images and multiple visual evidences in two images. A *Bayesian Belief Network* (BBN), also known as a *Bayesian Network*, is a graphical representation of a joint probability distribution over a set of random variables [10, 13]. A BBN is a directed acyclic graph in which each variable is represented by a node, and directed edges between nodes represent conditional dependencies. The dependencies can represent the causal influences among variables. Given a set of N variables $\mathbf{V} = \{V_1, \ldots, V_N\}$, the joint probability distribution $P(\mathbf{V})$ can be factored in any number of ways using Bayes' rule. A BBN exploits independencies between variables to specify the joint distribution over $\mathbf{V}$ via a sparse set of conditional probabilities: $P(\mathbf{V}) = \prod_{i=1}^{N} P(V_i|\Pi(i))$ where $\Pi(i)$ is the set of parent nodes of node i.

To perform inference, the user observes a subset **e** of $\mathbf{V}$, the N variables, referred to as *evidence*. After incorporating this evidence into the network, the distribution represented is $P(\mathbf{V}|\mathbf{e})$, that is the distribution of all variables given the available evidence. Note that not all variables need to be observed for inference to take place. Given the distribution $P(\mathbf{V}|\mathbf{e})$, marginalisation yields the distribution of each variable given the evidence, $P(V_i|\mathbf{e})$. Thus, the matching problem defined by Equation (1) can be probabilistically inferred by obtaining a probability distribution over the assignment combinations. If the network is a poly-tree, i.e. there is only one path connecting any two nodes, inference can be performed on the original network structure using the method introduced by Pearl [13]. If the network contains undirected cycles, this inference algorithm becomes intractable because the messages can cycle forever. The network must first undergo a series of transformations to obtain a *junction tree* in which inference can be performed using message passing [10].

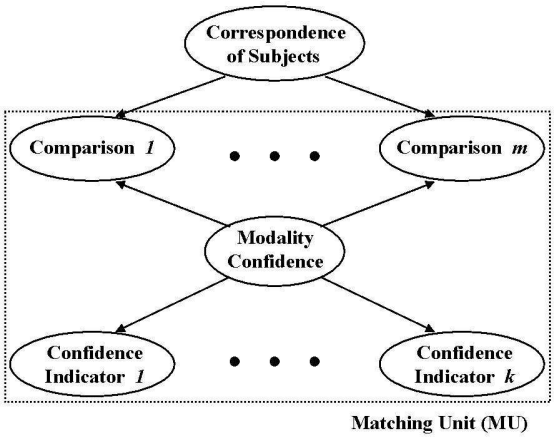

Figure 2. The Bayesian Network for inferring the correspondence of subjects between two images based on a single modality.

In the discrete-variable BBN (Figure 2) used to match subjects between two images based on a single modality, there are four different types of nodes: (1) Correspondence node which represents a multi-values variable and each value corresponds to a possible assignment combination $\{A_1, \cdots, A_{m!}\}$. The m is the maximum number of subjects in an image. (2) Comparison node. There are m comparison nodes and each node compares one subject in I_i against all m subjects in I_j. Thus, each subject in an image is compared to all subjects in the other image. (3) Modality confidence node which represents the confidence of the modality and constrains the influence of this modality on the correspondence. (4) Indicator node which indicates the modality confidence. Both correspondence and modality confidence nodes are the variables to be inferred. Comparison and indicator nodes are the variables representing the visual evidence. All the observed continuous values of observations are discretised. The conditional probability tables for the observation nodes can be learnt from a training set of data, either through statistical sampling in the case of complete data, or using the Expectation-Maximisation algorithm when some variables are unobservable [8]. From the observed evidence in indicator nodes, the modality confidence is inferred. This confidence and the computed comparison results are considered in inferring a probability dis-

tribution over the $m!$ assignment combinations. Note that the maximum number of subjects in an image is m. To obtain the distribution when the number of subjects in two images is less than m, the distribution can be marginalised from the inferred probability distribution over $m!$ assignment combinations. When the number of the subjects are different in two images say p and q in I_i and I_j with $p > q$, the $p-q$ subjects in I_j are replaced with null subjects. Thus, the less likely subjects in I_i will not be assigned any subjects in I_j. In order to generalise the BBN for multiple modalities, we define a Matching Unit (MU) as the union of all comparison, modality confidence and confidence indicator nodes.

3. Single Camera Tracking

To track people with a single camera, the system performs two major tasks: detecting the moving people and matching the subjects between consecutive frames. To take advantage of the fact that the camera is stationary, the moving subjects are segmented using a simple frame differencing method. After thresholding and noise cleaning, connected component analysis is applied to the foreground pixels to find the moving blobs, though it is not always correct. Each detected blob is then circumscribed by a bounding box and the system assigns an identity to this new detected blob.

To reliably maintain the identities of the detected people, the system integrates multiple modalities based on motion continuity and the apparent colour (Figure 3). A second-order Kalman filter is attached to each subject to estimate the motion vectors, $\mathbf{Z}_x(k) = [x, \dot{x}, \ddot{x}]^T$ and $\mathbf{Z}_y(k) = [y, \dot{y}, \ddot{y}]^T$, of the blob centroid. Furthermore, the colour data is sampled from the subject image and the distribution is modelled as Gaussian mixture models in hue and saturation space [18]. The conditional probability of a measured pixels, $\boldsymbol{\lambda}$, being the subject, $\mathcal{S}$, modelled as a mixture with m components is given by:

$$p(\lambda|\mathcal{S}) = \sum_{i=1}^{m} p(\lambda|i)P(i) \quad (2)$$

where $P(i)$ is the prior probability of the component, the i^{th} component is a Gaussian with mean μ and covariance matrix Σ, and:

$$p(\lambda|i) = \frac{1}{2\pi|\Sigma|^{1/2}} exp\{-\frac{1}{2}(\lambda-\mu)^T\Sigma^{-1}(\lambda-\mu)\} \quad (3)$$

The normal method to define the closest match based on Kalman filter is by searching the minimum of $\mathcal{M}_m = \nu^T \mathbf{S}^{-1} \nu$ where the ν is the innovation, the error between the predicted measurement and the true measurement, and $\mathbf{S}$ represents the covariance of the innovation. The closest match based on colour can be found as the blob with the minimum $\mathcal{M}_c = \sum_{j=1}^{n}\sum_{i=1}^{m}[(\lambda-\mu)^T\Sigma^{-1}(\lambda-\mu)]p(i)$ where n is the number of pixels sampled from the blob. These $\mathcal{M}_m$ and $\mathcal{M}_c$ are the *Mahalanobis Distance* (MD) for each individual blob used to quantify the likelihood and decide the match in the comparison node. Experiments performed show that the reliability of both modalities degrades when the features are not extracted accurately. In order to address this problem, we define the confidence indicators for motion as the status of the size of the detected blobs, the aspect ratio of the blob bounding box and the motion continuity based on the centroid displacement. Similarly, we use the status of the blobs as the confidence indicator of the colour. Moreover, we define $\sum_i \sum_{j\neq i} d_{i,j}$ where d is the distance between means, μ, of the dominant Gaussian of each subject in the subsequent frame. This value is used as another confidence indicator and can causes the system to rely less on colour information when the colours between subjects are similar.

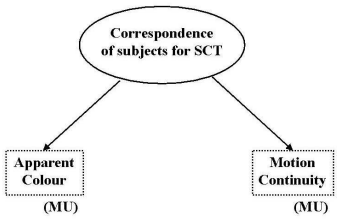

Figure 3. The general representation of Bayesian Network used to integrate multiple modalities for matching subjects between two consecutive frames in Single Camera Tracking (SCT). (MU see Figure 2)

After detection and identity assignment, the system tracks people in each single camera independently. When the status of segmented blobs change suddenly or the matching becomes ambiguous, the system performs MCCT (see Section 4) to pass identities between cameras. To determine the matching ambiguity, we apply the $\mathcal{X}^2$ test to the MD of each pair of the inferred assignment combination because the MD is $\mathcal{X}^2$ distributed [4]. In general, this statistical test has a corresponding critical value which defines the probability that a true match with MD larger than this value. We chose the critical value corresponding to 5% probability to decide the matching is ambiguous. Therefore, when any corresponding pair with MD of either modality greater than the critical value, the system will perform MCCT.

One important issue can not be ignored in tracking is the computational complexity of the correspondence problem [15]. In order to cope with this problem, different methods were proposed to reduce the number of candidate matches before matching is performed, such as small velocity change, smooth motion constraints and the bucket method [15]. This step is sometimes referred to as feature validation [4]. The $\mathcal{X}^2$ test we used to decide the matching ambiguity can also be used for feature validation.

4. Multi-Camera Cooperative Tracking

Our system performs SCT in each single camera continuously. Once the tracking becomes ambiguous in a camera, the system performs MCCT to resolve the ambiguity by matching subjects across cameras. To do this matching, system employs BBN to integrate five modalities as shown in Figure 4. In the following, we first introduce 3 geometry-based and 2 recognition-based modalities to be used for Bayesian modality fusion. Then, we describe how the system integrates multiple modalities.

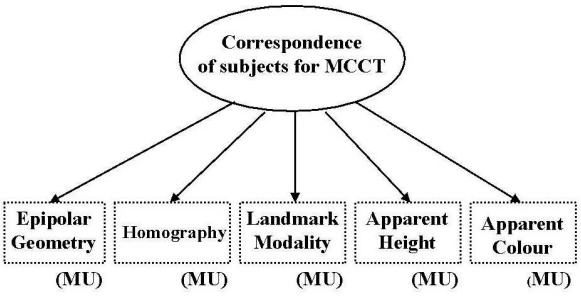

Figure 4. The general representation of Bayesian Networks to integrate multiple modalities for matching subjects across camera views in Multi-Camera Cooperative Tracking (MCCT).(MU see Figure 2)

4.1. Geometry-based modalities

For geometrically constraining the image positions of corresponding subjects, we use the multiple view geometry and a landmark method. The geometry of multiple views is well understood and has had a rapid increase in application to computer vision in last decade [7]. For two views, there exist constraints that relate the 3D corresponding points in two views to the camera geometry. Given a set of coplanar points, the constraints take the form of the homography. For a set of 3D general points, the constraints are the epipolar geometry.

Epipolar geometry: To apply epipolar geometry for matching, the topmost point of segmented blob in the first camera image I_1 is used to compute its associating epipolar line in the second camera image I_2. The distance between the epipolar line and topmost point of the subject in image I_2 is used as a match score. We assume that such a distance is a Gaussian variable with zero mean and a probability density function defined as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{x^2}{2\sigma^2}\} \quad (4)$$

The likelihood of the subject in I_2 being the corresponding subject in I_1 is determined by the value of the density function for the measured distance. We define $\mathcal{M}_e = \frac{x^2}{\sigma^2}$ and use it to compare the candidate matches. The modality confidence indicator is defined by the mean distance between affine epipolar lines. Moreover, we also use the segmentation status of the topmost point to indicate the confidence.

Homography: The general method to apply homography for tracking is to assume that people move on the ground plane and the bottom points lying on the planes are used to match subjects across camera views [12]. However, in some indoor environments, the lower part of subjects are not visible due to occlusion or being chopped by the lower view boundary. We use the topmost point of a person's head and assume this point lies on the same virtual plane when he/she is moving. Once a person is matched in two views, the topmost point pairs are used to estimate the homography for this particular person. Then for different people with different heights the system estimates different homographies to deal with different virtual planes.

To match the subjects between two camera images I_1 and I_2, we first transform the feature point (x, y) of a blob in I_1 to a point (x', y') in I_2. This projected point is then used to compute $\mathbf{x}' = (x', y', \dot{x'}, \dot{y'})$, called the *kinematic vector*, where $(\dot{x'}, \dot{y'})$ is the spatial displacement between consecutive frames of I_2. The matching is based on the comparison of this kinematic vector. We again apply a Gaussian variable with zero mean to model the difference between the projected kinematic vector and the observed kinematic vector, $\mathbf{x}$, of its corresponding subject in I_2. Thus, the matching likelihood of a subject in I_2 is given by:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^2|\Sigma|^{1/2}} exp\{-\frac{1}{2}[(\mathbf{x}-\mathbf{x}')^T\Sigma^{-1}(\mathbf{x}-\mathbf{x}')]\} \quad (5)$$

We define $\mathcal{M}_h = [(\mathbf{x}-\mathbf{x}')^T\Sigma^{-1}(\mathbf{x}-\mathbf{x}')]$ and use it to compare the candidate matches. We also define the confidence indicator for homography modality in terms of the segmentation status of the topmost point again, and mean distance between subjects' topmost points in I_2.

Landmark modality: Here, we utilise the scene knowledge based on multiple vertical line landmarks to constrain the image positions of the corresponding subjects. The vertical lines can be easily found from the man-made objects in an indoor environment. From this knowledge, the position of a subject with respect to the landmarks in an image , called Vertical Area (VA), can be used to constrain the positions of its corresponding subject in the other image (Figure 5). The modality confidence indicator is defined as the segmentation status of the topmost point used to determine the VA position of a subject. Moreover, the mean distance between the topmost point and the closest line landmarks is also used. This is because the VA position might not be reliable when a subject's topmost point is too close to the landmark due to wrong segmentation. However, geometric modalities alone do not provide enough constraints to match subject across cameras. In the next section, recognition-based modalities are described.

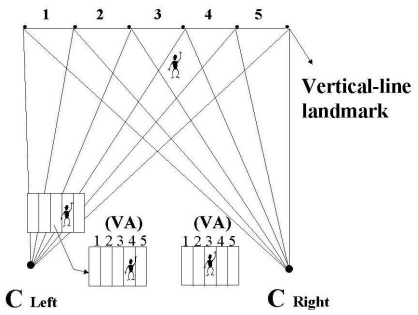

Figure 5. Top view of the FOVs of two cameras. From the Vertical Area (VA) position, n, of the subject in image 1, it can be inferred that the VA of its corresponding subject in image 2 is equal to or less than n, assuming people in the overlapping FOVs.

4.2. Recognition-based modalities

This matching method is based on the the similarity test of the subjects' image patterns. We use the apparent height and apparent colour of the subject. Since the appearance is view-variant, the system should estimate the appearance of the corresponding subject across camera views and use this "corrected" value for matching. To learn the mapping of the appearance, we fist partition the room into small virtual vertical volumes based on the vertical line landmarks and each volume is represented as (x_1, x_2) where x_i is the subject's VA position in image I_i. Then, we estimate this mapping by employing Support Vector Regression (SVR) [5] for each different small virtual vertical volumes in the world assuming the mapping is the same in each single volume. More detail about this mapping estimation and the landmark method discussed in previous paragraph can be found in [2].

Apparent height: The apparent height of a subject is defined as the longest distance in the vertical direction of a blob. This height is determined by a person's height and viewing geometry. Since our system is stationary, the correlation between the apparent height of a person in two views is fixed and can be used as a subject feature for matching. For a subject with VA x_1 in I_1, the apparent height of its corresponding subject in I_2 can be estimated from the learnt mapping. Since we do not know which is the corresponding subject in I_2, the system uses the mapping corresponding to the volume (x_1, x_2) to estimate the apparent height h' for each subject in I_2 according to its VA x_2. Again, we model this difference $h - h'$ as a Gaussian variable with zero mean. Thus, the matching likelihood of a subject in I_2 is given by:

$$f(h) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{(h-h')^2}{2\sigma^2}\} \qquad (6)$$

We define $\mathcal{M}_{ht} = \frac{(h-h')^2}{\sigma^2}$ and use it to compare the candidate matches. We defined the confidence indicator in terms of the segmentation status of the feature points used for computing apparent height and the mean difference of subjects' heights in I_2.

Apparent colour: As mentioned in SCT, apparent colour of a subject clothes image is modelled as Gaussian mixture models. Similar to the apparent height, for a subject in I_1 the apparent colour of its corresponding subject in I_2 can be estimated from the learnt mapping. The estimation is done for each single Gaussian model of the apparent colour. The mapping between colours in two views is learnt for different colours and this mapping can generalise to an "unseen" colour. Then, the match likelihood, similar to SCT, can be obtained based on the estimated colour models for subjects in I_2. The MD and confidence indicator for this modality is defined as the same as those used for colour modality in SCT.

4.3. Matching subjects across cameras

Having discussed the multiple modalities, we shall now describe details of the use of Bayesian modality fusion for matching subjects across cameras as shown in Figure 4. To fuse multiple modalities for matching, our system use the accumulated evidence in order to make the matching more reliable and smooth. We define $\sum_{i=0}^{q} \alpha\mathcal{M}(k-i)$ and use it to compare subjects where k is the frame index and α is the weight to set more recent evidence with higher weights. Note that if a modality is not reliable, the comparison will be based on the accumulated errors. In this case, the modality confidence indicator can adjust the influence lower.

To obtain consistency, the network is coupled indirectly over time through the specification of prior probability for correspondence node. As a consequence, the correspondence at each time instant is affected by the previous matching history. However, the matching might be incorrect when the visual information is not reliable. The system needs a method to prevent using the wrong information from the previous results. We apply $\mathcal{X}^2$ test, similar to SCT, to each pair of the assignment combination obtained from previous frame. If any pair with more than one modalities larger than the critical value, the system does not use the previous matching results in the correspondence node. Moreover, the number of frames of accumulated evidence used in comparison node is set as $q = 0$ for all modalities to prevent using wrong evidence. Thus, $\sum_{i=0}^{q} \alpha\mathcal{M}(k-i) = \mathcal{M}(k)$ and the system compares subjects based on the current frame images. Once the system continues to infer the same assignment combination, it stops performing MCCT and assigns the identities to the matched subjects.

Here, we discuss our method to reduce the complexity for MCCT based on the homography. Generally the bucket method or some other constrains, such as smooth motion constraints, do not apply to this problem without assumption. One possible method is to apply homography and assume ground plane is viewed in both views. In this case, the system can perform the bucket method on the ground plane. When the bottom points are not viewed, the topmost point can be used if the homography induced by the virtual plane

of each individual's topmost point has been estimated. First, the topmost points of all subjects in the unambiguous image I_1 are used to transform to ambiguous image I_2. Then, for each subject in I_2, compute $\mathcal{M}_h$ for all subjects in I_1 and apply the $\mathcal{X}^2$ test on these $\mathcal{M}_h$. Only the subjects in I_1 with $\mathcal{M}_h$ smaller than the critical value need to be considered for mating this subject in I_2. From this feature validation, the system can eliminate the less likely subjects before matching based on other modalities. Compared to other modalities homography is a powerful constraint which ideally can obtain a corresponding point across cameras. Another possible method to reduce the candidate matches is to use the domain knowledge such as our landmark method or the spatial relationship of the FOVs [9]. This knowledge based method can constrain the image positions of corresponding subjects in both views. The system can eliminate some less likely subjects before matching.

5. Results

Our multi-camera system was implemented on a SGI workstation with two uncalibrated cameras: a SGI digital camera and a SONY EVI-D31. The experiment is conducted by using these widely separated static cameras to monitor a room. In the following, a tracking example is used to demonstrate how the system match subjects across cameras in order to maintain the identity and solve the occlusion problem. Note that only part of the floor can be viewed by the cameras, so the homography based on ground plane does not apply. Since a large room is unavailable to test our theory of feature validation based on the homography related to the topmost point, the system considers all subjects in two views. We also demonstrate that Kalman filters follow the wrong people due to direction change during occlusion. Finally, we demonstrate the performance evaluation of our Bayesian modality fusion for matching subjects across cameras.

5.1. A tracking example

The tracking example, in Figure 6, consists of 450 frames with three people interacting with each other. In order to test our system, all three people are wearing red clothes such that the algorithm can not distinguish them based on colour alone. The bounding box corresponding to each blob is the segmented region based on background subtraction. The label on top of a bounding box is the identity assigned by the system when the person first appear in either view. The white cross is the topmost point of a blob. Figure 6 shows during the whole sequence, the system can maintain identities consistently based on Bayesian modality fusion even the occlusion is present in a view. To illustrate the working of our modality fusion approach, we highlight a section of this sequence beginning from when person 1 is in both views and person 2 just enters the room imaged by the right camera but not the left (Figure 6.a). As person 2 enters the left FOV, both people are in the overlapping FOVs (Figure 6.b) and the system performs MCCT to obtain the identity from the other camera assuming the subjects in two views correspond to the same people. From the topmost points of two subjects in the right view, I_2, the epipolar line (black) is used for searching subjects in the left, I_1. The topmost point of person 1 is also transformed to I_1 (black dot on top of person 1) based on the on-line learnt homography to compare with the observed kinematic vector of two subjects in I_1. The system can not use homography related to person 2 for since person 2 just enter the room and his/her related homography has not been estimated yet. It also can be seen that the topmost point of person 2 was incorrectly segmented. During matching, the epipolar geometry , colour and height modalities are less reliable, and the homography and landmark methods are dominant. Note that although the information is incomplete and less reliable, the BBN can still effectively collect evidence and make a right match.

After entering, person 2 continues to walk towards the room centre and person 1 towards the door. These subjects meet in I_1 and are segmented as one region, but not in I_2 (Figure 6.c). The system interprets that I_1 is ambiguous and relies on the tracking results from I_2 to disambiguate. The black dots in I_1 are the transformed points from the topmost points (white dots) of two subjects in I_2 based on its own stored estimated homography. From modality fusion, the merged blob in I_1 is matched to and interpreted as person 1 due to the top point of this blob corresponding to person 1.

Occlusion resolved by MCCT: Here, we demonstrate that the system can successfully maintain the identities after occlusion by using two cameras cooperatively where the Kalman filter may fail. When the merged blob splits into two blobs, the system detects that the number of blobs changes and performs MCCT as shown in Figure 6.d. After matching, the system passes the identifier of two people from the I_2 to I_1. Person 2 keeps walking to the right corner and person 1 turns and faces person 2. At this moment, another person enters the room and is assigned a new identity (Figure 6.e). Person 1 then turns around and walks toward person 3 (Figure 6.f). Similar to Figure 6.c, occlusion happens in I_1 as shown in Figure 6.g, but two people change direction during occlusion. To resolve the occlusion, the homography is more reliable than the other modalities. It can be seen that in Figure 6.h, the transformed points in I_1 can be reliably used to search for corresponding people and the epipolar geometry is less reliable.

Kalman filter failure: Note that tracking with a single camera can correctly resolve the ambiguity in the event of Figure 6.c, but can not maintain correct identities for the event Figure 6.g. Figure 7 illustrates the tracking failure with a single camera based on motion continuity for the latter event. It shows the measured (ground truth) and the predicted positions of the blob centroids of person 1 and 3 in I_1. During occlusion, the position estimation is based on a constant velocity assumption and the acceleration is not used because it is unreliable. The Kalman filters can follow people before occlusion, but fail to estimate correct positions of people after occlusion.

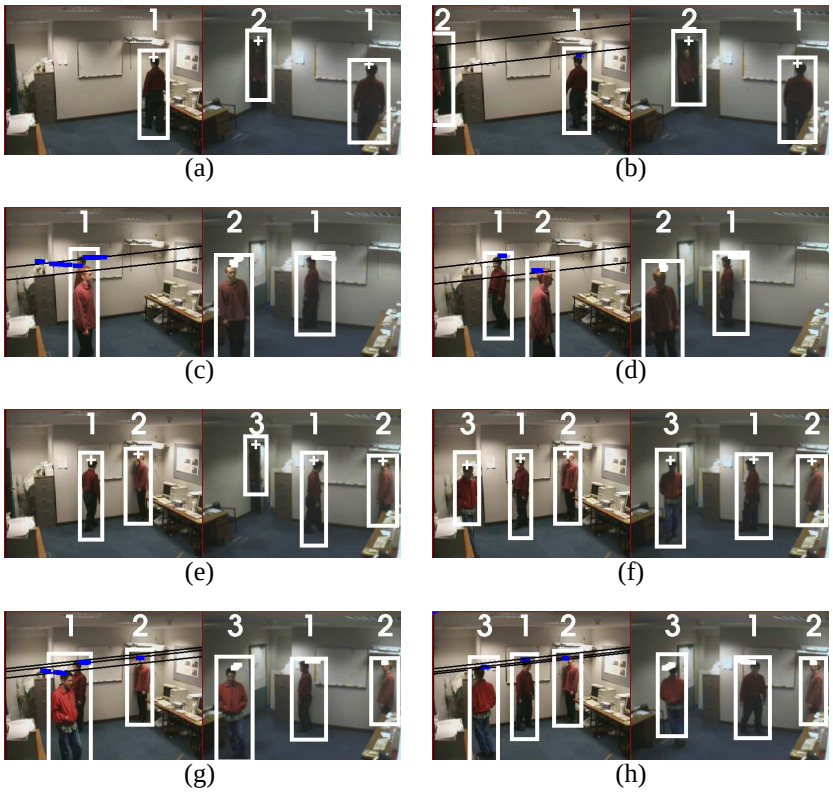

Figure 6. The system can track people with identities using two cameras cooperatively even when occlusion is present.

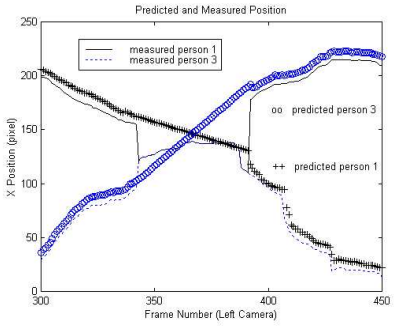

Figure 7. The measured (ground truth) and predicted blob centroids of person 1 and 3 in the left view of the tracking example (Figure 6). Occlusion is present at frame 343-395 as shown in Figure 6.(f-h). The Kalman filter fails to estimate the positions after occlusion due to change direction during occlusion.

5.2. Performance evaluation

To highlight the strength of Bayesian modality fusion for combining multiple cues, we compare it with a popular fusion method for tracking. This method assumes all modalities are independent, often called the *naive Bayes*, and the match result is given by $M(\mathbf{S}, \mathbf{S}') = \prod_{k=1}^{n} P(a_k|a'_k)$ where $\mathbf{S}$ and $\mathbf{S}'$ represents two subjects to be matched with n different features a_k and a'_k respectively. In order to compare the robustness of these two methods, we collected 20 sequences of two people interacting with each other in the overlapping FOVs. The people had a wide range of heights, colour of clothes and various motions and the sequences were captured under various lighting conditions. Figure 8 illustrates the results of matching two people between two camera views. The accuracy rate of each sequence is the overall matching accuracy of all frames. The ground truth of matching is generated by hand. The average accuracy of all 20 sequences is about 99.1% with deviation 1.2% for the Bayesian modality fusion and 96.5 % with deviation 2.4% for the naive Bayes method. We found that using BBN is better in combining multiple visual evidences for matching subjects across cameras.

6. Discussion

We now discuss the strength and weakness of our system and future improvement directions. We have demonstrated

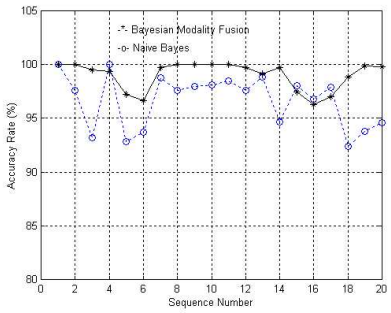

Figure 8. The accuracy rate of matching subjects between two camera views based on Bayesian modality fusion and a naive Bayes method for 20 image sequences.

that our multi-camera tracking system can handle occlusion and maintain identities of multiple people. Handling occlusion using appearance and motion is in general hard because the image pattern of the subject appearance can experience severe variation during occlusion and the motion model might be violated during the estimating stage (as in Figure 7). The former situation is inherently difficult because there is no strong temporal model to predict the appearance. For example, the colour constancy problem in our tracking example can result in a wrong match and it might be more serious in an environment with multiple illuminants. Moreover, the limited indoor space confines the location of the camera which can in turn prevent using some domain knowledge to resolve the matching ambiguity such as ground plane constraints or a world model of the tracking environment.

To apply homography related to the topmost point for matching subjects across cameras is restricted to people with upright pose. At this moment, we use this method when the image position of this point does not change suddenly assuming the person is in the same pose and the top point of his/her head keeps lying on the same virtual plane. We would like to be able to recognise the people's poses in order to apply this method more reliably. Another limitation is that the position of the camera must be high enough such that the homography does not degenerate as the plane projected as a line.

From our experiments, we also found that wrong segmentation, such as shadow, causes the system to fail. This problem can cause the system match wrong people. It also can cause the system to miss some people or cause a false alarm. This problem can be alleviated by using multiple cameras. For example, the false target can be "deactivated" by using multiple cameras since the shadow might not be imaged in the other cameras. This is one of the advantages of using a multi-camera system: the system has more chances of obtaining unambiguous information.

References

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

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

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

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

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

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

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

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

[9] O. Javed, S. Khan, Z. Rasheed, and M. Shah. Camera handoff: Tracking in multiple uncalibrated stationary cameras. In *IEEE Workshop on Human Motion*, TX, USA, 2000.

[10] F. V. Jensen. *An Introduction to Bayesian Networks*. UCL Press, 1996.

[11] S. Khan and M. Shah. Tracking people in presence of occlusion. In *Asian Conference on Computer Vision*, 2000.

[12] L. Lee, R. Romano, and G. Stein. Monitoring activities from multiple video streams: Establishing a common coordinate frame. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Special Issue on Video Surveillance and Monitoring:758–767, 2000.

[13] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[14] P. Pritchette and A. Zisserman. Wide baseline stereo matching. In *ICCV*, pages 863–869, Bombay, India, 1998.

[15] K. Rangarajan and M. Shah. Establishing motion correspondence. *CVGIP: Image Understanding*, 54:56–73, 1991.

[16] R. Rosales and S. Sclaroff. Improved tracking of multiple humans with trajectory prediction and occlusion modeling. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1998.

[17] K. Toyama and E. Horvitz. Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *Asian Conference on Computer Vision*, Tapei, Taiwan, January 2000.

[18] R. Yogesh, S. J. Mckenna, and S. Gong. Segmentation and tracking using colour mixture models. In *Asian Conference on Computer Vision*, pages 607–614, Hong Kong, 1998.