BAYESIAN MODALITY FUSION FOR TRACKING MULTIPLE PEOPLE WITH A MULTI-CAMERA SYSTEM

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

Shaogang Gong
Department of Computer Science
Queen Mary, University of London, London E1 4NS, UK
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. A Bayesian network is used to combine multiple modalities for matching subjects between multiple camera views. Unlike other occlusion reasoning methods, we use multiple cameras 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

Human tracking in an indoor environment is of interest in a number of applications such as visual surveillance and human-computer interface. 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 by using multiple uncalibrated static and widely-separated cameras.

Different solutions to the occlusion problem in human tracking have been proposed. Rosales and Sclaroff [1] used Kalman filters and Khan and Shah [2] used colour. However, neither of these methods work

Figure 1. The task is to track people with identities even under occlusion using two widely separated cameras.

for all cases. Recently, Haritaoglu et al. [3] 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 system needs to pass the subjects identities across cameras once the identities are lost in a certain view by matching subjects across camera views. To this end, Collins et al. [4] use the trajectory and normalised colour histogram of an object. Chang et al. [5] estimate the subjects' apparent height and apparent colour across cameras. This matching can also be done by geometric method, such as epipolar geometry [6], homography [9] and landmarks [5]. However, these feature-based matching methods can be unreliable due to the ambiguous positions of the extracted features resulting in inconsistencies. A framework is required to combine multiple visual modalities, or cues, to make the matching more reliable. 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 Single Camera Tracking (SCT), the system then passes identity and assigns it to this subject by matching subjects across camera views, called Multiple Camera Cooperative Tracking (MCCT).

To track people with a single camera, the system first performs frame differencing. After thresholding and noise cleaning, connected component analysis is applied to the foreground pixels to find the moving blobs. To reliably maintain the identities of the detected people, the system integrates multiple modalities based on motion continuity and the apparent colour. A second-order Kalman filter is attached to each subject to estimate the motion of the blob centroid. Furthermore, the colour of the subject image is modelled as Gaussian mixture models in hue

and saturation space. The conditional probability of a measured pixels, $\boldsymbol{\lambda}$, being the subject, $\mathcal{S}$, modelled as a mixture with m components is given by: $p(\boldsymbol{\lambda}|\mathcal{S}) = \sum_{i=1}^{m} p(\boldsymbol{\lambda}|i)P(i)$ where $P(i)$ is the prior probability of the component, the i^{th} component is a Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and:

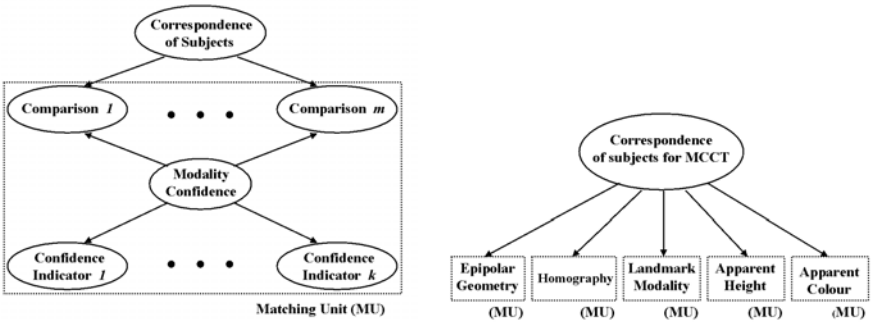
$$p(\boldsymbol{\lambda}|i) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} exp\{-\frac{1}{2}(\boldsymbol{\lambda}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\lambda}-\boldsymbol{\mu})\} \quad (1)$$

The 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 and $\mathbf{S}$ represents the covariance of the innovation. The closest match based on colour can be found as the blob with the minimum $\mathcal{M}_c = \sum_{j=1}^{n}\sum_{i=1}^{m}[(\boldsymbol{\lambda}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\lambda}-\boldsymbol{\mu})]p(i)$ 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. When the matching becomes ambiguous, determined by applying $\mathcal{X}^2$ test to the MD of each pair of the match [7], the system performs MCCT to pass identities between cameras.

2. Bayesian Networks for Building Correspondence

To define matching problem for MCCT, we firstly constrain the maximum number of subjects in each image to be m. To match subjects between 2 camera images, I_i and I_j, we evaluate the matching globally, i.e. consider the matching for all subjects simultaneously. In each combination of assignment, every subject in I_i is assigned a corresponding subject in I_j. After applying the uniqueness constraint, there could be $m!$ possible assignment combinations, $A_\alpha = \{A_1, \cdots, A_{m!}\}$. Given the visual evidence $\mathbf{e}$ from cameras, our goal is to find a most appropriate assignment combination which maximises the posterior $p(A_\alpha|\mathbf{e})$.

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 [8]. The defined matching problem can be probabilistically inferred by obtaining a probability distribution over the assignment combinations. In the discrete-variable BBN (Figure 2(a)), 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 $\{A_1, \cdots, A_{m!}\}$. (2) Comparison node. There are m comparison nodes



(a) (b)

Figure 2. (a)The Bayesian Network for inferring the correspondence of subjects between two camera images based on a single modality. (b)The general representation of Bayesian Networks to fuse multiple modalities for matching subjects across camera views.

and each node compares one subject in I_i against all m subjects in I_j. (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. 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. Multi-Camera Cooperative Tracking

In the following, we first introduce 3 geometry-based and 2 recognition-based modalities for MCCT and then describe how the system integrates multiple modalities.

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:

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

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: 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. 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. 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: From the scene knowledge based on the vertical-line landmarks, 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. 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.

The recognition-based modalities are based on the the similarity. Since the appearance is view-variant, the system should estimate the appearance across camera views and use this "corrected" value for matching. We employ Support Vector Regression (SVR) for estimation. More detail can be found in [5].

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. Again, we model this difference $h - h'$ between the estimated height and the observed height as a Gaussian variable with zero mean. 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.

To fuse multiple modalities for matching with the BBN (Figure 2), our system use the accumulated evidence $\sum_{i=0}^{q} \alpha \mathcal{M}(k-i)$ to compare subjects where k is the frame index and α is the weight to set more recent evidence with higher weights. 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. We apply $\mathcal{X}^2$ test, similar to SCT, to each pair of the assignment combination obtained from previous frame. If any pair fails the test, 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. Once the system continues to infer the same assignment combination, it stops performing MCCT and assigns the identities to the matched subjects.

4. Results

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. To illustrate the modality fusion approach, we highlight a section of the sequence beginning from when person 1 is in both

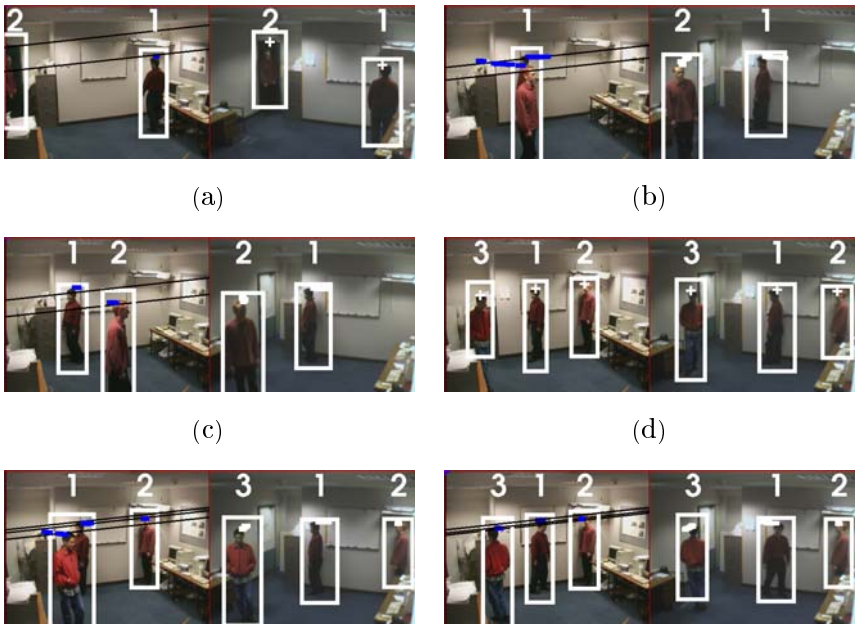
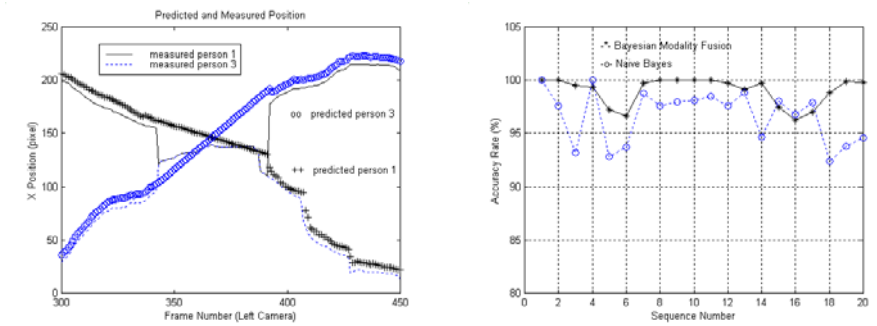


Figure 3. The system can track people with identities using two cameras cooperatively even occlusion is present.

views and person 2 just enters the room already imaged by the right camera, I_2, and first appearing in the left I_1 (Figure 3.a). The system performs MCCT to obtain the identity for this new detected subject from the right camera. From the topmost points of two subjects in I_2, the epipolar line (black) is used for searching subjects in 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. It also can be seen that the topmost point of person 2 I_2 in was incorrectly segmented, but the BBN can still effectively collect evidence and make a right match.

After entering, person 2 continues to walk towards the room centre and these two subjects meet in I_1. The system interprets that I_1 is ambiguous and relies on 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. When the merged blob splits into two blobs, the system detects that the number

of blobs changes and performs MCCT as shown in Figure 3.c and passes the identifier from I_2 to I_1. Then person 3 enters the room, occlusion happens in I_1 as shown in Figure 3.e, but two people change direction during occlusion. Note that tracking with a single camera can correctly resolve the ambiguity in the event of Figure 3.b, but can not maintain correct identities for the event Figure 3.d-f. Figure 4(a) illustrates the tracking failure with a single camera based on motion continuity for the latter event. The Kalman filters can follow people before occlusion, but fail to estimate correct positions of people after occlusion.



(a) (b)

Figure 4. (a) The measured (ground truth) and predicted (Kalman filter) blob centroids of person 1 and 3 during occlusion in the left view of the tracking example (Figure 3(d-f)). The Kalman filter fails due to change direction during occlusion. (b)The accuracy rate of matching subjects between two camera views based on two methods.

To highlight the strength of Bayesian modality fusion we compare it with a popular fusion method assuming all modalities are independent, often called the *naive Bayes*. Figure 4(b) 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 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 more accurate in combining multiple visual evidences for matching subjects across cameras.

5. Discussion

We have demonstrated 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. From our experiments, we also found that wrong segmentation, such as shadow, causes the system to fail. This problem can be alleviated by using detection results of multiple cameras. This is the advantage of using a multi-camera system: more chances of obtaining unambiguous information.

References

[1] R. Rosales and S. Sclaroff. (1998). *Improved Tracking of Multiple Humans with Trajectory Prediction and Occlusion Modeling.* IEEE International Conference on Computer Vision and Pattern Recognition.

[2] S. Khan and M. Shah. (2000). *Tracking People in Presence of Occlusion.* Asian Conference on Computer Vision.

[3] I. Haritaoglu, D. Harwood and L. Davis. (1998). *W4: Who? When? Where? What? A Real Time System for Detecting and Tracking People.* IEEE International Conference on Automatic Face and Gesture Recognition.

[4] R. Collins, A. Lipton, T. Kanadeand, H. Fujiyoshi, D. Duggins and Y. Tsin. (2000). *A System for Video Surveillance and Monitoring: VSAM Final Report.* Carnegie Mellon University. CMU-RI-TR-00-12.

[5] T. H. Chang, S. Gong and E. J. Ong. (2000). *Tracking Multiple People Under Occlusion Using Multiple Cameras.* British Machine Vision Conference.

[6] Q. Cai and J. K. Aggarwal. (1998). *Automatic Tracking of Human Motion in Indoor Scenes Across Multiple Synchronized Video Streams.* IEEE International Conference on Computer Vision.

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

[8] F. V. Jensen. (1996). *An Introduction to Bayesian Networks.* UCL Press.

[9] L. Lee, R. Romano and G. Stein. (2000). *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.