A Framework for Track Matching Across Disjoint Cameras using Robust Shape and Appearance Features

C. Madden and M. Piccardi
Faculty of Information Technology
University of Technology, Sydney
Sydney, NSW 2007

Abstract

This paper presents a framework based on robust shape and appearance features for matching the various tracks generated by a single individual moving within a surveillance system. Each track is first automatically analysed in order to detect and remove the frames affected by large segmentation errors and drastic changes in illumination. The object's features computed over the remaining frames prove more robust and capable of supporting correct matching of tracks even in the case of significantly disjointed camera views. The shape and appearance features used include a height estimate as well as illumination-tolerant colour representation of the individual's global colours and the colours of the upper and lower portions of clothing. The results of a test from a real surveillance system show that the combination of these four features can provide a probability of matching as high as 91 percent with 5 percent probability of false alarms under views which have significantly differing illumination levels and suffer from significant segmentation errors in as many as 1 in 4 frames.

1. Introduction

Computer vision-based object tracking is based upon shape, motion, and appearance features. Motion features have tended to be widely utilised in human environments, such as within buildings, because of the previously limited camera resolution to exploit shape or appearance features effectively. Particular instances of human motion do not necessarily conform to statistical expectations, especially in existing surveillance system which consist of a limited number of cameras that are disjoint, often significantly, around the surveillance area. This makes motion information unreliable for estimating how a person may move when they are not observed. While true that coverage can be improved by using many overlapping or near-overlapping cameras for important areas, it proves too expensive to be the general case. Fortunately, the increasing resolution of cameras is providing more and more accurate shape and appearance information that can support effective tracking even with sparse cameras. This is already assisting security officers to monitor and follow suspects of interest throughout a system such as what occurred in the 2005 London bombings. However, much precious time could have been saved had an inter-camera, automated tracking system existed. Creating such a system faces a range of difficulties including differing camera properties and viewpoint, imperfect object segmentation, occlusions, and variable and unpredictable illumination conditions.

Previous work in matching the tracks of an individual across disjoint cameras has focused upon colour matching [3, 6, 8, 9] as a key feature for matching people, though some biometrics such as gait and height estimates have also been explored for this purpose [1, 3, 10]. We herewith clarify that in this work we use the word 'track' to refer to the information obtained from the uninterrupted tracking of a single individual. Such information includes the indexes of the starting and last frame of the track, the blob's region and appearance in each frame and features that can be derived from them such as gait and shape. Individuals can very often be discriminated based on the colours of their clothing, with the notable exception of people wearing uniforms; however illumination changes pose a major problem to the invariance of appearance features. Javed *et al.* [6] proposed to compensate illumination variations by training their system to recognise sets of frequent illumination conditions in order to transform colours to a normalised colour set. However, this approach cannot compensate for different illumination in different regions of the images. Gandhi and Trivedi [8] present a cylindrical representation of the individual to obtain spatial colour information using multiple overlapping cameras. This is likely to be sensitive to the alignment of the cylindrical representation and articulated motion, even where overlapping cameras are available. Darrel *et al.* [3] propose the fusion of facial patterns with height and colour features. Unfortunately, typical surveillance cameras do not offer sufficient illumination for accurately measuring facial patterns. BenAbdekader *et al.* [1] present height estimation based upon converting the height of an individual object's bounding box into a measurement using camera properties without a full camera calibration. Stride length and periodicity are also determined; however, they require a frame rate higher than twice the gait frequency and reasonably stable which may not always occur in standard surveillance systems. Model-based gait features such as those developed by Zhou *et al.* [15] may also prove useful as additional features for the track matching framework presented in this paper.

Individuals analysed for tracking systems are segmented using many different techniques depending upon the complexity of the scene, the computational speed requirements of the application, and the type and availability of colour and stereo cameras. We segment individuals using a background subtraction method similar to Pfinder [13] to provide fast segmentation, even though other widely used approaches [12] may generate less errors. A degree of errors occur in all current techniques [11], and could significantly affect any shape and appearance features. However to date we are unaware of any existing methods that try to identify frames in a track where segmentation is poor, as opposed to selecting which are the most reliable segmentation techniques for the given scene. Therefore, we have developed and included a technique for identifying such errors in order to extract robust shape and appearance features to use in the proposed track matching framework.

2. Track Matching Overview

The presented work is based upon the definition of the surveillance session as a portion of one day as people enter the surveillance area from a known entry point to perform their activities before leaving through a known exit point. This definition leads to the following simplifying assumptions about the surveillance area, and the people viewed within it:

- All entry and exit points of the surveilled area are in view of a surveillance camera.
- Individuals are unlikely to change their clothing or footwear; hence, many of their intrinsic shape and appearance features will remain relatively constant for the duration of the session.
- Individuals are tracked accurately whilst within the view of any of the system's cameras.
- Individuals are segmented from the background into a single blob, sometimes merged with other objects.
- Individuals are generally observed at a distance from the camera.
- Where cameras are spaced apart by sizeable distances, motion features may vary unpredictably between those cameras as individuals are allowed free motion.

The above assumptions suggest that shape and appearance features should be used rather than motion features to generate accurate matching of object tracks within any surveillance system where camera views do not overlap. Unfortunately, due to the articulated motion of people, few shape features other than height or gait are likely to remain stable during walking. In addition, appearance features relating to clothing are likely to remain stable within the extent of a surveillance session. For these reasons, we have based our track matching framework upon extraction and comparison of upper clothing, lower clothing, and global colour appearance as the appearance features outlined in Section 3. Height was chosen as a reliable shape feature as outlined in Section 4. Gait is currently not exploited for our surveillance system due to its slow and variable frame rate. While exceptions may be easily constructed, this feature set is designed to provide sufficient discrimination (at the ground truth level) for a large majority of real cases.

The appearance features are analysed for changes along the track that indicate significant segmentation errors as described in Section 5. This allows the system to automatically extract the reliable frames from an individual's track sequence to generate more robust features for the track matching process. These robust shape and appearance features are then compared between any two tracks to compute a track similarity figure for each feature. A small training set is then used in Section 6 to determine likelihood functions for matching and nonmatching cases conditioned to the similarity values. Bayes theorem is then applied to fuse the likelihoods of the similarities to determine if any two tracks are matching.

3. Extracting and Comparing Appearance Features

The Major Colour Representation (**MCR**) used in this paper to define the colour features extends the method previously developed in [2]. We propose to add two extra colour features relating to the upper and lower clothing colours of an individual to the global colours used in [2, 3, 5]. These features are chosen to represent the often different colours of the clothing on the upper torso, and those on the legs. The narrow spatial aspect of these features also allows for a more sensitive analysis of the spatial positioning of a persons colours. This ensures that changes in the position of the colours can be detected, such as where segmentation errors remove large portions of the object, and used to discriminate between people wearing similar colours on different portions of their body.

Extracting the MCR for each of these three colour features utilises the same process, but analyses different spatial component of the appearance of the segmented object. The process for extracting the MCR's is described in detail in

[2], but summarises as:

- A controlled equilisation step performs a data-dependent intensity transformation that spreads the histogram to compensate for some degree of illumination change that can be expected within the indoor and outdoor surveillance environments.
- Online K-means clustering of pixels of similar colour within a normalised colour distance generates the MCR of each spatial region.
- Once segmentation errors are removed, robust MCR features can be obtained over a small window of frames to improve robustness to articulated motion.

The three colour features are:

- The global MCR feature representes the colours of the whole segmented object without any spatial information.
- The upper MCR feature represents the colour of the top portion of clothing. This corresponds to the region from $30 \div 40$ percent of the person from the top of the objects bounding box as shown in Figure 1. This narrow band was chosen to ensure that it avoids the inclusion of the head and hair of the object, as well as low necklines, but does not go so low to overlap with the leg area.
- The lower MCR feature is aimed to represent the colour of the lower portion of clothing. This corresponds to the region from $65 \div 80$ percent of the object from the top of the objects bounding box as shown in Figure 1. This narrow band avoids the very bottom of the object which can be prone to shadows, or artifacts where the feet touch the ground. It also tries to avoid overlapping with the belt or upper torso area of the person.

The narrowness and positioning of both of the upper and lower MCR regions also allows for them to remain constant under minor segmentation errors that will only have a minimal impact upon a person's features, whilst still remaining sensitive to large segmentation errors. These features also allow for the inclusion of spatial colour features which improves the identification of differences between individuals when tracking is incorrect.

Figure 1 show the upper MCR feature region (enclosed between the two top lines) and the lower MCR feature region (between the two bottom lines). Positioning of such regions in the first and third frames in figure is regarded as acceptable. In the second frame, instead, the lower MCR feature region is significantly displaced; in this case, the sudden change in the lower MCR feature clearly indicates a very poorly segmented frame.

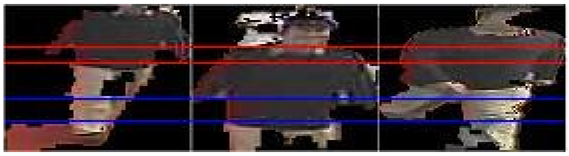

Figure 1: Example of upper and lower regions from three segmentations of an individual

4. Extracting and Comparing Shape Features

In this section, we describe the method proposed to estimate the height of a walking person and the height difference between two people from any two disjoint tracks. In [10] we showed that height estimation can be achieved using the single camera views that dominate surveillance systems, although a camera calibration step needs to be performed, and the individual needs to be fully segmented with reasonable accuracy for these measuments to be useful. Here, we extend [10] by automatically extracting the key positions at the top of the head and a reasonably accurate estimation of the ground plane position. The pairwise height differences between each of the frames from two tracks can be statistically analysed to determine the similarity as:

$$s_H = \frac{\sigma(Hd)}{\mu(Hd)} \tag{1}$$

where $\mu(Hd)$ and $\sigma(Hd)$ are the average and standard deviation estimate of *Hd*, respectively.

The following steps outline the height difference estimation process:

1. Determine the height estimate of the objects in each frame of the track:
   (a) The silhouette of the object is analysed using a *k*-curvature technique [7].
   (b) Areas near the bottom of the object with high curvature *k* are then used to determine where the feet are positioned, and thus extract a midpoint at the bottom of the object *b(u,v)*.
   (c) This point is converted into world ground plane coordinates *b(x,y,z)* to determine location.
   (d) This location is then used with the image plane position of the top of the head *h(u,v)* to estimate the height of the person, *Hf*.
2. The pairwise differences, *Hd*, between each frame and every other frame are computed between the tracks.
3. The estimated height differences between the object tracks are statistically analysed to determine s_H in **(1)**.

4.1 Automatically determining the top and bottom object position

We propose to locate the position of the top of the head and the feet position as precisely as possible from a monocular view. The feet position is found using a *k*-curvature technique [7] after the object has been segmented from the background. The *k*-curvature technique follows the chain of silhouette pixels determining the curvature at each pixel based on three pixels along the curvilinear coordinate, x_1, x_2 and x_3:

$$k = tan^{-1}\left(\frac{x_1 - x_2}{y_1 - y_2}\right) - tan^{-1}\left(\frac{x_3 - x_2}{y_3 - y_2}\right) \quad (2)$$

If $k < 0$ then we use $k = 2\pi + k$ to ensure $0 < k < 2\pi$.

When areas of high *k*-curvature occur in the bottom 30 percent of the object, they are likely to correspond with an individual's feet, or more particularly their toes or heels. If the position of the two points is relatively close together around the silhouette, then they are set to correspond to the heel and toe of one foot of the object, and can be averaged to produce a foot position estimate. Otherwise, the single significant curved area is used as the foot position estimate. Where two areas are found with high curvature in the lower portion of the object, but are relatively far apart of the silhouette, then they are assumed to represent the two separate feet, and are analysed accordingly. The two feet estimates can then be found and averaged to provide an estimate of a midpoint at the bottom of the object $b(u, v)$ as shown in Figure 2, which we use as the ground plane position $b(x, y, z)$. The usage of a bottom point tends to reduce the gait effects on the determination of the ground plane position as it reflects the centre of balance of the person as they walk. It also tends to produce a better height estimate than simply using the middle of the bottom of the bounding box.

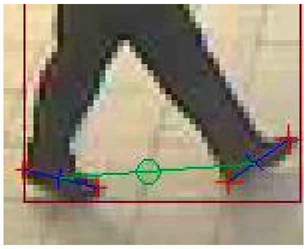

Figure 2: Finding $b(u, v)$ using two feet

The head top $h(u, v)$ position is calculated much more simply from the object silhouette as we assume that a person is standing vertically, which is the most common case in surveillance areas. It uses the midpoint of the top row of object pixels as the middle of the top of the head, which is different to simply using the midpoint of the top of the bounding box. The automatically extracted head *h(u,v)* and bottom positions *b(u,v)* can then be converted from the top left image plane coordinate system into real world ground plane coordinates using camera calibration as shown in [10]. This produces an estimate of the height of the segmented object within the frame. The differences in height of all the frames in track A from track B are defined as the set of height differences *Hd* and can statistically determine the similarity measure using **(1)**.

5. Identifying Segmentation Errors

Identifying major segmentation errors is the first step required to extract robust shape and appearance features to be later used for a number of subsequent functions. We propose to do this by analysing the changes in the features of a segmented individual along the frame sequence in which they were tracked. Our method assumes that individuals were tracked accurately using one of the many popular tracking systems, such as [14]. We also assume that segmentation in the majority of the frames is affected only by minor errors (as a pre-condition for successful tracking) and we aim to detect those frames where major segmentation errors occurred due to transient occlusion, cluttering, or major lighting changes.

For this purpose, we utilise the extracted global, upper and lower MCR colour features and we compared them for each frame pair in the track. These features share the same colour representation and same comparison technique [9] outlined in Section 3. This comparison produces a similarity value for each of the three MCR features for any frame pair. A statistical analysis over a known training set of frame pairs with major changes *H0* and without *H1* provided us with likelihood functions conditioned to such similarities. For any unseen frame pair we can then determine the change *H0* or non-change *H1* hypothesis based on its similarity values and the likelihoods. We assume the features to be independent even if they are not completely and so apply Bayes theorem as:

$$P\left(H0|s_G, s_U, s_L\right) = B\left(P\left(s_G|H0\right)P\left(s_U|H0\right)P\left(s_L|H0\right)\right) \quad (3)$$

$$P\left(H1|s_G, s_U, s_L\right) = P\left(s_G|H1\right)P\left(s_U|H1\right)P\left(s_L|H1\right) \quad (4)$$

where B is a prior that can be used to bias the operating point of the system, and the maximum of 3 and 4 resolves the classification.

Where the majority of frames are designated as not matching the current frame, or *H0*, then the frame most likely has segmentation errors large enough to distort any shape or appearance features. By using this process, we can automatically identify and remove such frames from the computation of the object features to be used later for track matching.

6. Fusing Matching and Non-matching Robust Object Feature

When comparing two tracks, we compute a track similarity value for each of the features (this should not be confused with the similarity values described above which are computed between two frames). Here, we utilise Bayes theorem again to fuse together the track similarities from each feature as shown in **(5,6)**, where s_H is the height similarity, s_{UC} relates to the upper clothing colour, s_{LC} relates to the lower clothing colour and s_{GC} relates to the global colour. This method also allows for the extension of the feature framework by adding extra terms to **(5,6)** relating to the *H0* and *H1* in a similar manner.

$$P\left(H0|s_H, s_{UC}, s_{LC}, s_{GC}\right) = B\left(P\left(s_H|H0\right)P\left(s_{UC}|H0\right)P\left(s_{LC}|H0\right)P\left(s_{GC}|H0\right)\right) \quad (5)$$

$$P\left(H1|s_H, s_{UC}, s_{LC}, s_{GC}\right) = P\left(s_H|H1\right)P\left(s_{GC}|H1\right)P\left(s_{LC}|H1\right)P\left(s_{GC}|H1\right) \quad (6)$$

The classification of the track pair is simply provided by the maximum probability between hypotheses *H0* or *H1*. The fusion scheme would hold for additional features provided they can be treated as mutually independent.

7. Results

The results presented here report track matching accuracy based on each separate feature and for the fused case based upon a comparison of 26 tracks from four people across two cameras, giving over 300 possible comparison combinations. Of these, 60 comparison combinations are used as training data with the remaining used for testing. An indication of the clothing's colour and good segmentation examples for the four individuals is given in Figure 3, where it is easy to see that the individuals are wearing clothing of approximately 50 percent or more differing colours. Ground-truth height differences between the individuals range from approximately 5 centimeters to 30 centimeters. We present the accuracy of the given feature components, which can then be compared with the fused results. The results indicate that the features complement each other to reduce the rate of false matches, which can be seen in the ROC curves of the independent and fused variables shown in Figure 4.

Figure 4 demonstrates that the fusion of the chosen features can provide a probability of detection of 91 percent with only 5 percent false alarms at our chosen operating point. The accuracy of the major segmentation error detection was as high as 84 percent with only 3 percent false alarms, indicating that whilst most of the erroneous frames were identified and discarded, the vast majority of reliable frames remained available. Our results also showed that some cameras may be likely to produce more accurate segmentation and feature measurements than others due to increased contrast between the individual and the background, and more stable lighting conditions. This would indicate that the feature probability distributions could possibly be better defined based upon the camera pairs within which the tracks occur; however this was not utilised for these experiments. Compensation of the effects of variable illumination on the object's appearance was performed as proposed in [9]; however if colour calibration of the camera was applied, then it might further improve the discriminative power of the colour features. The maximum-a-posteriori classification of **(5)** and **(6)** based upon the knowledge that only 1 in 4 tracks are matching can minimise the total Bayesian error. However, we prefer to work at a different operating point along the ROC curve in order to achieve a higher probability of detection, because a human operator can quickly and easily identify a falsely matched pair by doing a fast manual track review as the difference in appearance is likely to be reasonably obvious. Manually correcting a missed detection is more arduous as an operator would then need to manually compare the current track to

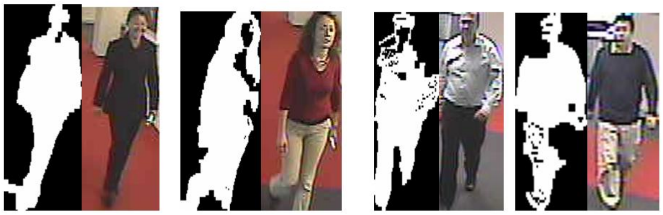

Figure 3: Four people of interest and good automatically segmented masks

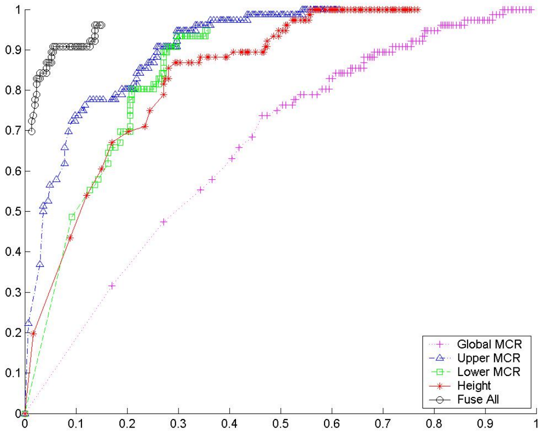

Figure 4: ROC curves of the height, colour and fused feature results to be revised

all other possible tracks to determine the best match. Hence, we have opted to adjust B in **(5)** by a factor of three, achieving the results reported. As already stated this method relies upon good segmentation; however the overall detection rates show that this method works well with automated detection and removal of frames with segmentation errors.

8. Conclusion

This paper has presented a framework to fuse robust shape and appearance features of individuals so they can be matched when observed in subsequent cameras. The robustness of the features is achieved by identifying frames within track of an individual where significant segmentation errors occur. A system was implemented based upon this framework using height as a shape feature and three appearance features relating to the individual's global colours, as well as upper and lower clothing colours. The results of this system using footage obtain from a real surveillance system achieved major segmentation error detection rates as high as 84 percent, with only 3 percent false alarm, allowing the retention of most reliable frames, even when errors occurred in as many as 1 in 4 frames. The results of track matching indicates that the careful choice of spatial colour appearance features and height as a shape feature complement each other to provide a high level of accuracy when they are fused together, even when frame rates are low and object segmentation is often poor. A detection rate of almost 91 percent with 5 percent false alarms indicates that such a system could be useful as an automated method to identify the movements of key inidividuals throughout a surveillance system. With a manual revision of the combination of matched tracks, very high levels of accuracy could be achieved to enhance forensic investigations of the movement of individuals throughout a surveillance system.

Acknowledgments

This research is supported by the Australian Research Council under the ARC Discovery Project Grant Scheme 2004 - DP0452657.

References

[1] C. BenAbdekader and R. Cultler and L. Davis, "Person Idnetificaiton using Automatic Height and Stride Estimation," *International Conference on Image Processing*, pp. 377-380, 2002.

[2] E. Cheng, C. Madden, M. Piccardi, "Mitigating the Effects of Variable Illumination for Tracking across Disjoint Camera Views," *International Conference on Advanced Video and Signal Based Surveillance*, pp. 32-38, 2006.

[3] T. Darrell and G.Gordon and M. Harveille and J. Woodfill, "Integrated Person Tracking Using Stereo, Colour, and Pattern Detection," *International Journal of Computer Vision*, Vol. 37, No. 2, pp. 175-178, 2000.

[4] A. Elgammal and R. Duraiswami and L.S. Davis, "Efficient kernel density estimation using the fast gauss transform with applications to color modeling and tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 11, pp. 1499-1504, 2003.

[5] C. E. Erdem and F. Ernst and A. Redert and E. Hendriks, "Temporal Stabilization of Video Object Segmentation for 3D-TV Applications," *International Conference on Image Processing*, pp. 357-360, 2004.

[6] O. Javed and K. Shafique and M. Shah, "Appearance Modeling for Tracking in Multiple Non-Overlapping Cameras," *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 26-33, 2005.

[7] H. Freeman and L. Davis, "A corner-finding algorithm for chain-coded curves," *IEEE Transactions on Computing*, Vol. 26, pp 297-303, 1977.

[8] T. Ganhdi and M. Trivedhi, "Panoramic Appearance Map (PAM) for Multi-Camera Based Person Re-Identification," *Advanced Video and Signal Based Surveillance*, 2006.

[9] C. Madden and E. D. Cheng and M. Piccardi, "Tracking People across Disjoint Camera Views by an Illumination-Tolerant Appearance Representation," *Machine Vision and Applications*, 2007.

[10] C. Madden and M. Piccardi, "Height Measurement as a Session-based Biometric for People Matching Across Disjoint Camera Views," Image and Vision Computing New Zealand, pp. 282-286, 2005.

[11] R. J. Radke and S. Andra and O. Al-Kofahi and B. Roysam, "Image Change Detection Algorithms: A Systematic Survey," *IEEE Transactions on Image Processing*, Vol. 14, No. 3, pp. 294-307, 2005.

[12] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 246-252, 1999.

[13] C. Wren and A. Azarbayejani and T. Darrell and A. P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Transactions on Pattern Analysis And Machine Intelligence*, Vol. 19, No. 7, pp. 780-785, 1997.

[14] T. Zhao and R. Nevatia, "Tracking Multiple Humans in Complex Situations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 9, pp. 1208-1221, 2004.

[15] Z. Zhou and A. Prugel-Bennet and D. R. I. Damper, "A Bayesian Framework for Extracting Human Gait Using Strong Prior Knowledge," *IEEE Transactions on Pattern Analysis And Machine Intelligence*, Vol. 28, No. 11, pp. 1738-1752, 2006.