Identifying Customer Behaviour and Dwell Time using Soft Biometrics

Simon Denman, Alina Bialkowski, Clinton Fookes, Sridha Sridharan

Abstract In a commercial environment, it is advantageous to know how long it takes customers to move between different regions, how long they spend in each region, and where they are likely to go as they move from one location to another. Presently, these measures can only be determined manually, or through the use of hardware tags (i.e. RFID). Soft biometrics are characteristics that can be used to describe, but not uniquely identify an individual. They include traits such as height, weight, gender, hair, skin and clothing colour. Unlike traditional biometrics, soft biometrics can be acquired by surveillance cameras at range without any user co-operation. While these traits cannot provide robust authentication, they can be used to provide identification at long range, and aid in object tracking and detection in disjoint camera networks. In this chapter we propose using colour, height and luggage soft biometrics to determine operational statistics relating to how people move through a space. A novel average soft biometric is used to locate people who look distinct, and these people are then detected at various locations within a disjoint camera network to gradually obtain operational statistics.

Simon Denman
Image and Video Laboratory, Queensland University of Technology, Brisbane, Australia, e-mail: s.denman@qut.edu.au

Alina Bialkowski
Image and Video Laboratory, Queensland University of Technology, Brisbane, Australia, e-mail: alina.bialkowski@qut.edu.au

Clinton Fookes
Image and Video Laboratory, Queensland University of Technology, Brisbane, Australia, e-mail: c.fookes@qut.edu.au

Sridha Sridharan
Image and Video Laboratory, Queensland University of Technology, Brisbane, Australia, e-mail: s.sridharan@qut.edu.au

1 Introduction

For operators of large, public facilities such as airports, transport hubs and shopping centres, it is important to understand how people move through the environment and how long it takes to move between key points. At present, only manual methods are available to collect such data, typically achieved by a staff member handing a customer a piece of time stamped paper which they hand to a second staff member at a later point in the process.

While object tracking systems [20, 6, 54, 17, 18] offer a possible solution for small environments and situations where there is continuous camera coverage, they are not practical for deployment in locations such as an airport, where there are large and frequent gaps in the camera network and a large number of possible paths through the environment (i.e. it is unclear which camera a person would next appear in). While techniques such as crowd counting [55] are designed to operate in such an environment, they only extract information about the crowd as a whole, and disregard the behaviour of individuals within it.

Within a large, disjoint, surveillance environment, soft biometrics [15, 12, 53] offer a means to continuously recognise people as they move through an environment. However, in a crowded space where there are hundreds, or possibly thousands of people present, there are likely to be a large number of people who have a similar appearance making accurate matching across views challenging and error prone. For example, in a typical airport there are likely to be a large number of business men wearing dark suits. However, while it would be difficult to match these similar looking people, a person who stands out from the crowd would be comparatively easy to follow through the environment, even when the path they take and speed which they move through the environment could not be anticipated.

In this chapter, we propose using soft biometrics to model a persons appearance, and from observations of multiple people calculate an average soft biometric. By comparing people to the average biometric, we can then determine how unique their appearance is, allowing us to automatically select only individuals who stand out from the crowd. This sub-set of people can then be re-detected throughout a disjoint camera network using their soft biometrics (see Figure 1), to obtain operational measures such as the time taken to travel from point to point.

We demonstrate the proposed system on a small database of up to four cameras collected in-house, and show that the average soft biometrics can be used to locate individuals with a distinct appearance, and then match these people across disjoint camera views. We show that this can be used to obtain an accurate estimate of the time it takes for people to move through the environment.

The remainder of this paper is structured as follows: Section 2 presents an overview of soft biometrics in surveillance imagery and other related literature; Section 3 outlines the soft biometric models used in this work; Section 4 describes how the average biometrics are calculated and how these are used to detect distinct individuals and calculate operational statistics; Section 5 presents and evaluation of both the soft biometric models (Section 5.1) and the estimation of operational measures

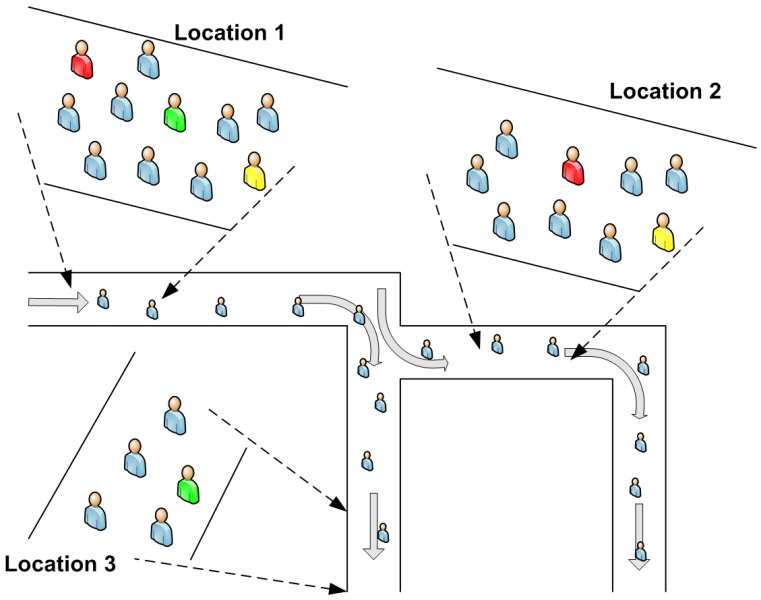


Fig. 1 Detecting distinct people - At location 1, three distinct looking people can be detected (yellow, green, and red). As the people move through the environment, two of these distinct people can be re-detected at location 2 (yellow and red), and the third can be re-detected at location 3 (green).

through the detection of unique individuals (Section 5.2); and Section 6 concludes the chapter and outlines possible directions for future research.

2 Background

Soft biometrics are traits that describe a person, but cannot uniquely identify them. They include information such as gender, ethnicity, height and build, and clothing colour. To date, soft biometrics have had two main uses: as a means to improve the performance of traditional biometrics systems by incorporating soft biometrics [33, 1, 44, 46]; or as a way to recognise people in surveillance footage [15, 12, 53].

The initial uses of soft biometrics focused on improving the accuracy of existing 'hard' biometric systems. Jain et al. [33] proposed the incorporation of gender, ethnicity, and height information of a person in fingerprint and face recognition using a Bayesian framework for fusion; Ailisto et al. [1] combined body weight and fat measurements to improve the performance of fingerprint biometrics; Marcialis et al. [44] used hair colour and ethnicity to complement face as the primary biometric; while Park et al. [47] proposed incorporating gender and facial marks (i.e. scars, freckles, moles, etc.) to improve face recognition. Niinuma et al. [46] proposed using soft biometrics for continuous authentication of a user at a computer, combining

face and clothing colour with traditional face recognition and the users password. In all situations, soft biometrics were able to improve the recognition performance of the combined system.

In addition to identification improvements, soft biometrics can be used to more quickly and efficiently filter a large biometric database. For example, Wayman [65] proposed the use of gender and age for filtering a large biometric database. This limits the number of entries in a database to be searched, and hence significantly reduces the time required for identification. However, if a person is incorrectly classified, recognition performance may be degraded.

Although these systems have successfully used soft biometrics, the focus is on complementing primary biometric systems such as fingerprint and face. In such situations, the soft biometric features are trained and classified on high resolution still images, which are well aligned and often captured under controlled lighting conditions. For soft biometrics in video surveillance, the extraction of soft biometric features is a more challenging task due to lower quality data and uncontrolled environment, such as changes in illumination, resolution, the pose of a person, camera view angle, as well as the presence of occlusions and shadows. In such conditions, extraction of some soft biometrics used in these systems, such as facial marks, is likely to be extremely difficult.

In a surveillance environment, traits that can be observed at a greater range are desirable, as are traits that are invariant to view and to lighting conditions. Cognitive experiments examining suitable traits for describing individuals have shown that individuals recognise other people by assigning them into categories including the soft biometric traits of gender, age, race and build or physique [43]. Samangooei et al. [56] quantified how accurate and useful human-provided descriptions of people are for recognising a person at a distance. Physical traits were categorised into semantic terms for humans to label video sequences of different people, and race and gender were found to be the most effective measures. Recent research has attempted automatic extraction of such traits from video. For example, the extraction of gender from gait [14, 69, 59], the estimation of age from gait [41], and determining ethnicity from faces [42]. While all these techniques show promise, they are difficult to apply in an unconstrained surveillance environment, and on their own do not yield sufficiently unique descriptions when dealing with a large population.

Appearance modelling techniques used in object tracking and person re-detection systems can also be viewed as a form of soft biometric. Appearance models are typically designed to be view and illumination invariant so that they may be used to aid in tracking handover between different camera views, and to aid in tracking during or after occlusions. Haritaoglu et al. [29] proposed a method where data pertaining to the average texture and silhouette of the subject is recorded over a period of time as the object is tracked. This model can be used to re-detect a person if they had been lost for several frames due to occlusion, or had left and re-entered the scene.

Whilst the approach proposed in [29] is effective for a single view, texture information less suitable for transferring from one view to another. Hu et al. [31] extracted three histograms from each person, one each for the head, torso and legs, to

not only allow for matching based on colour, but also on the distribution of colour. Chien et al. [7] proposed a colour model (Human Colour Structure Descriptor - HCSD) that aims to capture the distribution of colours in a human body. Three colours are used to represent the colour of the body, legs and shoes, and positions are defined to describe the position of body and legs relative to the shoes.

Nakajima et al. [45] and Hahnel et al. [25] each proposed techniques to model and recognise people based on their whole body. Nakajima et al. [45] extracts normalised colour histograms and local shape features for detected people and trains SVM classifiers for each person and pose, and the approach is shown to be accurate on a limited data set. Hahnel et al. [25] extends on [45] by extracting additional colour, shape and texture features.

To better evaluate appearance modelling and person re-detection techniques, Gray et al. [24] proposed the VIPeR (Viewpoint invariant Pedestrian Recognition) database, consisting of 632 image pairs, where each pair contains wide and varying changes in viewpoint, illumination and pose. Using the proposed data set, an evaluation of a variety of existing techniques such as 1D and 3D histograms, 3D correllograms and various template matching approaches was performed, and it was shown that a simple 1D histogram outperforms the more complex techniques.

Many recent approaches have sought to combine colour and texture features, and extract texture features which are less view dependant. Farenzena et al. [19] proposed an appearance-based method for person re-identification using symmetry-based features consisting of the overall chromatic content, the spatial arrangement of colours into stable regions (through the extraction of MSCRs [21]), and recurrent local motifs with high entropy (i.e. recurring textures). Symmetry is used to build the model through the use of weighted colour histograms computed along the symmetric axes, and by the sampling of patches to locate local motifs along the axes of symmetry; while the axes of asymmetry are used to segment the person into head, torso and legs. Bak et al. [2] proposed appearance models based on Haar-like features and dominant colour descriptors. The most invariant and discriminative signature was extracted using the AdaBoost algorithm.

Bazzani et al. [4] proposed a person descriptor that incorporates a global feature, in the form of a HSV histogram, and local features, determined through epitomic analysis [34]. A generic epitome (approximately equivalent to an 'average' texture of the person), and a set of local epitome (a set of recurring motifs within the subject) are extracted to represent the texture. As with the global feature, the epitomes are represented by a HSV histogram.

Schwartz et al. [57] proposed using a co-occurance matrix to extract a dense texture representation, as well as extracting edge and colour features for subjects. Features are extracted for a set of overlapping blocks within the subject region, leading to a very large initial feature vector. Dimension reduction is carried out using partial least squares (PLS), and a PLS model is learned for each subject in the gallery by using all other subjects as counter examples. This approach results in a model that highlights the most discriminative features for each subject. When classifying a probe image, the probe is projected into the subspace of each gallery subject, and the minimum Euclidean distance between the probe and gallery vectors

is used to classify the subject. The PLS scheme is shown to outperform a PCA and SVM approaches.

Other approaches have sought to improve person classification by incorporating the structure of the human body into the model. Bak et al. [3] proposed using body part detectors to locate individual body parts (head and shoulders, torso, left arm, right arm, legs) as well as the full body, and extracted features for each region. The covariance of features extracted from each region (location, intensity, and gradient) is used to match subjects within a spatial pyramid matching framework.

Xiaogang et al. [67] also sought to model the relationship between different parts of the human body and proposed the idea and shape and appearance context, which models the appearance of an object relative to its own components. From the input image shape and appearance labelled images are generated. The shape image describes the structure of the person (i.e. identifying body regions such as arms and legs), while the appearance image quantises the input image into a set of labelled regions. The co-occurance relationship between features within the different regions is extracted and is used to describe the appearance of a person.

Zheng et al. [71] extended the idea of appearance context to groups of people, observing that people often move in groups, and the appearance of the group provides additional cues to a person's identity. Two ratio-occurance measures for comparing groups are developed, one using rectangular rings around the image, and one using features within small image blocks, and it is shown that this combination of feature is effective at recognising groups of people.

Group context is also applied to an individual within the group, by extracting features in rings around the person of interest (providing a group context for the individual), in combination with features for the individual. It is shown that incorporating group context results in a significant performance improvement when identifying individuals.

Gheissari et al. [22] propose a multi-part person model, segmenting the person into 6 horizontal stripes, based on a decomposable triangulated graph model. A feature vector, composed of HSV colour information and edgels (where each edge encodes the edge orientation, and the colour change across the edge), is extracted for each of the horizontal stripes. To compensate for the dynamic properties of clothing which can result in edges changing rapidly, a spatio-temporal graph based segmentation approach to suppress edges is applied. This results in edges that occur within an item of clothing being ignored, while retaining those that relate to the boundary between different clothing items.

An alternative approach to crafting features specific to modelling an individual is to extract very high dimensionality features, and use a classifier to determine the optimal combination of these features to separate the classes. Gray et al. [23] proposed learning a similarity function using a combination of colour and texture (obtained by using the response to Schmid and Gabor filters) information. Localised features in horizontal stripes are computed for each possible feature (i.e. colour channel, or texture representation) and adaboost is used to learn the optimal combination of these features. The proposed approach is evaluated on the VIPeR data set [24] and is

shown to outperform the algorithms of Park et al. [48] and a principle axis histogram motivated by [32], as well as simple histogram and template representations.

Prosser et al. [51] and Zheng et al.[72] focused on finding an optimal ranking strategy to improve identification performance. Prosser et al. [51] formulated an alternative ranking strategy that does not rely on the direct computation of distances between gallery and probe models. A modification to the Rank SVM is proposed (termed an Ensemble Rank SVM) that aims to reduce the computational burden of Rank SVM by using a group of 'weak rankers'. A set of 'weak rankers' are learned on subsets of the overall data set, and the weak classifiers are recombined using boosting. The resultant system performs similarly to the traditional Rank SVM, with a significant reduction in computations requirements, using a set of features inspired by [23]. The proposed system also outperforms the ELF model of [23], as well as similar features matched using the Bhattacharyya and L1-norm distance measures.

Zheng et al. [72] proposed the probabilistic relative distance comparison (PRDC), that aims to learn the distance metric that maximises accuracy, to improve person re-identifiation. The proposed model is formulated to maximise the probability that the correct match for a given image will have a smaller distance to the probe image than an incorrect match (unlike traditional distance learning which seeks to minimise intra-class and maximise inter-class variation). Like other trained classifiers [23, 51], a large suite of features is used, and a set similar to [23] is chosen. The proposed approach is shown to outperform the learned classifier techniques for person re-identification of [23, 51], and is also shown to be more robust when the size of the training set is small.

An alternate matching strategy is proposed by Lin et al. [39], who outlined a pairwise matching approach, where the distance between each pair of gallery subjects is learned, and the match profiles that describe the similarity of the pairs are stored. For a given probe image, the match profile to all subjects in the gallery is computed and the matching subject is determined by comparing the set of profiles. Lin et al. [39] model appearance using normalised kernel density estimation, exploiting the non-parametric nature of kernel density estimation to model the complex colour distributions that people exhibit.

While a large amount of effort has been made to model colour and appearance, they are limited by changes in view. From different angles, people have vastly different appearances, and variations in the appearance of colours, both within a single camera and across multiple cameras within an environment, present further challenges. While some techniques such as [19, 4] aim to improve view invariance by extracting representative and recurring textures (rather than a dense texture representation), if the distance between views is sufficiently great (i.e. viewing the subject from the front, and then the back), these textures still may not re-occur and matching will fail. With this in mind, a variety of other features have been proposed to match subjects across multiple cameras.

The presence of luggage can be used as a descriptor, and a suite of techniques have been proposed to locate and model carried luggage. Approaches for detecting luggage or carried items include using contour analysis and SVM classification [52], locating regions of asymmetry and an analysis of the periodicity of these regions

[28], and through an analysis of simple motion region statistics in the case of [11], who sought to determine if a person is walking with a bicycle. Damen et al. [10] proposed a silhouette based technique that extends [28], comparing a time-averaged silhouette with a set of exemplars rather than assuming that a person is symmetric. Regions that protrude from the exemplar are identified, and constraints on the likely location of carried objects are enforced using a Markov random field. The resultant system is able to identify a wide variety of carried items, although it does rely on the carried object protruding from the person.

Various shape and size features have also been proposed to model people in tracking systems. The silhouette of a person obtained from motion detection is a popular shape features and can be used in a variety of ways. Collins et al. [8] has proposed the use of silhouette-based human identification from body shape and gait, and other features such as boundary distance from the centroid [64], and the convex hull can also be extracted.

Gait recognition techniques such as the Gait Energy Image (GEI) [26, 66, 70] can also be thought of as a soft biometric. The GEI is the average silhouette taken over a single gait period, enabling the temporal information of gait to be encoded in a single frame. Unlike gait techniques that rely on model-fitting, the GEI can easily be captured at long range in the same manner as other soft biometrics.

From an unconstrained surveillance standpoint, shape and gait based methods are limited due to their dependency on the viewing angle, and the quality of the extracted silhouette. Segmentation errors caused by complex background conditions and shadows can affect the performance. Whilst these errors can be overcome be averaging over a gait cycle provided that the errors are random and not systematic (shadows are not random and are still a large problem in complex lighting conditions), but errors caused by view mismatch are much more difficult to overcome and are likely to lead to recognition errors.

With the various limitations of the individual features, several approaches have been proposed that utilise multiple soft biometric and appearance features. Denman et al. [15] proposed modelling colour (as a three part head, torso and legs model) and height to describe people and identify them between camera views in a surveillance environment. Dantcheva et al. [12] described a weight modality, as well as a probabilistic colour histogram (also three part) that could be used to identify people in surveillance imagery. Ran et al. [53] proposed a gait signature, consisting of several soft biometrics based on gait features. Stride length, height and gender could all be extracted from a video sequence. Demirkus et al. [13] presented a system in which different biometric features of people are extracted based on the quality of the current frame. For example, in a frame where the whole body of a subject is visible, height and clothing colour features are extracted, while in another frame only the upper body might be visible, but the face may be quite clear and hence face-based features are extracted. All approaches have shown promise for use in recognition within a surveillance environment.

Soft biometrics have also been applied to the task of a visual search, i.e. locating a person in surveillance footage given a description. Park et al. [48] proposed extracting dominant colours, height and build (determined from the silhouette as-

pect ratio) to represent a subject. A query could then be submitted to the system to locate a person matching a description. Vaquero et al. [62] proposed an attribute based search, to locate people in surveillance imagery. Various facial features were extracted such as facial hair (beard, mustache, no facial hair), the presence of eye wear (glasses, sunglasses, no glasses) and headwear (hair, hat, bald), as well as full body features such as the colour of the torso and legs. Queries could be formulated as a combination of these features.

Soft biometrics and person re-detection techniques also have applications in customer behaviour analysis, and a variety of computer vision techniques and systems have been proposed to automatically analyse customer behaviour in retail environments. Many techniques have focused on counting people, with solutions proposed to count both everyone in an area [55, 38], and to count the number of people past a point (i.e. entering a shop) [35, 5]. Senior et al. [58] proposed a system that goes beyond simple counting, and developed a video analytics system for retail which counts the number of customers entering a store, and monitors where they go within the store.

Other vision systems have been developed to determine the interactions between people and products. Haritaoglu et al. [30] developed a body pose detection system using stereo cameras to automatically detect reaching actions of shoppers. Krahnstoever et al. [36] detected head and hand locations with stereo vision to automatically detect the interaction between the shopper and product, and RFID tags were used to track the location and motion of products.

Dwell time in front of products and billboards has also been explored. Haritaoglu et al. [27] developed a system to detect, track and count the number of people standing in front of billboards and extract their gender. Infrared illumination-based pupil detection was used to determine whether people were looking at the billboards, and for how long. Liu et al. [40] used Active Appearance Models (AAMs) to monitor the face for better accuracy in determining customer gaze.

Popa et al. [49] describes a surveillance system which analyses the shopping behaviour of customers in relation to products. A Bayesian Network architecture is employed that consists of three levels: sensor level, observed features level, and the customer's shopping behaviour (i.e. the assigned semantic label to the customer's actions). Through the use of motion detection, trajectory analysis, face localisation and tracking eight behaviours could be described: walking direction, walking speed, stopping, looking at products, facial expressions (for positive/negative appreciation), gestures, and speech. Experimental results on a data set of 10 subjects shows that by clustering the sensed features, and in particular the motion energy and walking trajectories, the approach can distinguish between customers behaviours as being either 'goal oriented' or 'disoriented'.

Popa et al. [50] also described a facial recognition system which analyses a customer's appreciation of a product based on their expression. Active Appearance Models are used to track the face and extract key features, and a Hidden Markov Models is used to classify the expressions into 21 product emotions, obtaining a recognition accuracy of 93%. By automatically detecting and analysing a customer's

expression, retailers can better fit products to the customers' needs, enabling more efficient marketing strategies.

These systems provide useful information on customer behaviour on a small scale, but do not scale to large environments. In large environments it may not be feasible to have cameras to monitor every space, and may not provide sufficient resolution for unique identification and so robust algorithms need to be developed to model individuals to enable tracking of people across disjoint camera views.

3 Soft Biometrics

Soft biometrics are features that can be easily extracted from a distance. Ideally, for use in an unconstrained surveillance environment any features should also be view invariant. In this work, we consider colour, height and luggage models for a person. We select these modalities as we consider them to be moderately view invariant and easy to extract in difficult conditions (i.e. colour and height can be determined from any angle, while texture, hair and skin features cannot be extracted consistently as the view changes). Furthermore, despite not being as discriminative as other recent approaches (for example [19, 3, 71]), both colour and height modalities are well suited to locating people who have an unusual aspect to their appearance, for instance people who are unusually tall or short, or who are dressed strangely. It should be noted however that other modalities (such as ethnicity, or more advanced colour and texture features) could also be used within the proposed framework.

The colour model is a three part model (head, torso and legs), and each section is modelled separately (see Section 3.1). The height of a person can be extracted using camera calibration, and is outlined in Section 3.2. Luggage can be detected by analysing the symmetry of the silhouette and locating regions of asymmetry. The detection and modelling of luggage is outlined in Section 3.3.

For all soft biometrics outlined within this Section, it is expected that the person of interest has been located within the scene (i.e. through object detection [9, 63] and/or object tracking [17, 6]), and a motion segmentation algorithm [16] is used to separate the subject from the background.

3.1 Colour

A 3D colour soft histogram is computed for each of the head, torso and leg sections (C_{head}, C_{torso} and C_{legs} respectively). The colour and motion image are used to generate histograms, such that only pixels that are in motion (i.e. part of the person) are included in the histogram. In order to separate the person into these three components, it is assumed that the person has been correctly segmented within the motion image (i.e. the motion regions relevant to the target person have been identified and

all other motion is disregarded), and that the person is predominately vertical within the image.

Segmentation of the person into head, torso and legs is performed in two stages:

1. Analysis of the horizontal projection to determine likely regions for the neck and waist;
2. Analysis of the gradient of the regions of interest to locate the neck and waste contour.

The horizontal projection,

$$P_{horiz}(j) = \sum_{i=0}^{W} M(i,j), \tag{1}$$

where $P_{horiz}(j)$ is the horizontal projection for row j, W is the image width, and M is the motion image; is computed for the region containing the person, and it is used to estimate the likely neck and waist positions. Given correct segmentation of the person in the motion image, the location of the neck will coincide with an increase in the horizontal projection (v_{neck}), and the waist (v_{waist}) will correspond with a gradual decrease in the projection. Figure 2 shows an example of this.

Typically, the neck and waist correspond with a change in appearance, and the a local maximum in the image gradient should be observed at this transition. A search region of 10% of the person's height is established around each of these points to determine the actual neck and waist boundaries (see Figure 2, (c) and (g)). Within each search region, each column is searched in turn for the location with the maximum vertical gradient. The resultant vector describes the contour that separates the two regions,

$$V_b(i) = \arg\max_{n=v_1}^{v_2} \delta I(i,n), \tag{2}$$

where V_b is the contour for the boundary b; v_1 and v_2 are the search bounds for the contour, set to 10% of the person's height either side of v_{neck} and v_{waist}; and $\delta I(i,n)$ is the gradient of the image I at location i,n. The computed contour is smoothed using a mean filter. Contours are computed for the neck and waist (V_{neck} and V_{waist} respectively), and these are used to divide the regions. Example output from this process is shown in Figure 3. This segmentation process assumes that the region being segmented is a single person, and the approach will produce unexpected results if a region consisting of multiple people is used as input. Within the proposed system, it is the responsibility of a prior process that locates the objects (such as an object detector or object tracker) to ensure that this segmentation is correct.

For each of the three regions, a 3D soft histogram is calculated to represent the appearance. Due to variations in colour across the different cameras, as well as possible errors in segmentation, there are a number of sources of error when computing and comparing histograms. In an effort to minimise the impact of these errors, we use a soft histogram. Soft histograms have previously been applied in a variety of areas, including activity modelling [68] and texture representation [37].

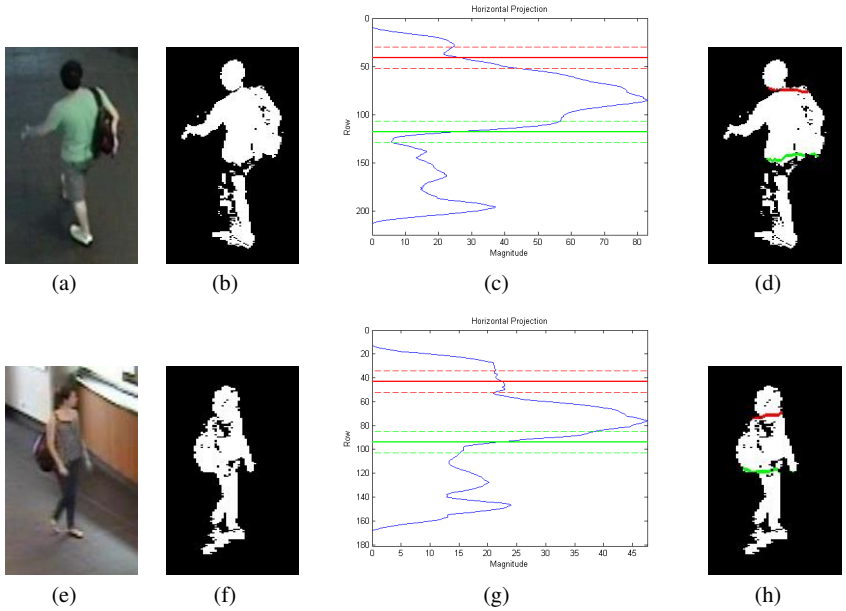


Fig. 2 Locating neck and waist boundaries - The top and bottom rows show two examples of extracting the neck and waist boundaries. Images (a) and (e) are the input colour images and (b) and (f) are the silhouettes computed through foreground segmentation. (c) and (g) show the smoothed horizontal projection, with the approximate neck (v_{neck}, red) and waist (v_{waist}, green) locations, and the search space for the boundary is shown by the dotted lines. (d) and (h) show the final boundary for the neck (red) and waist (green).

The soft histogram divides the contribution of each sample (pixel) across multiple bins. Weights are assigned based on the proximity of the sample to the centre of the bins. A sample that lies at the centre of a bin will have the majority of its contribution assigned to that bin, while a sample that lies on a bin boundary will have its contribution divided across multiple bins. For simplicity, we only consider neighbouring bins within the same dimension (i.e. we do not consider diagonal neighbours). This means that for a three dimensional histogram (i.e. where the bin a sample belongs to is a function of three colour channels: red, green and blue), a single pixel will have its weight divided across seven bins: the centre bin, and the two neighbouring bins in each dimension. The distance a sample is from each bin, and the corresponding weight, is calculated as follows,

$$d_{i,n} = \frac{3}{2} B_w - |b_n - i|\,, \tag{3}$$

$$w_{i,n} = \frac{d_{i,n}}{\sum_{m=1}^{M} d_{i,m}}, \tag{4}$$

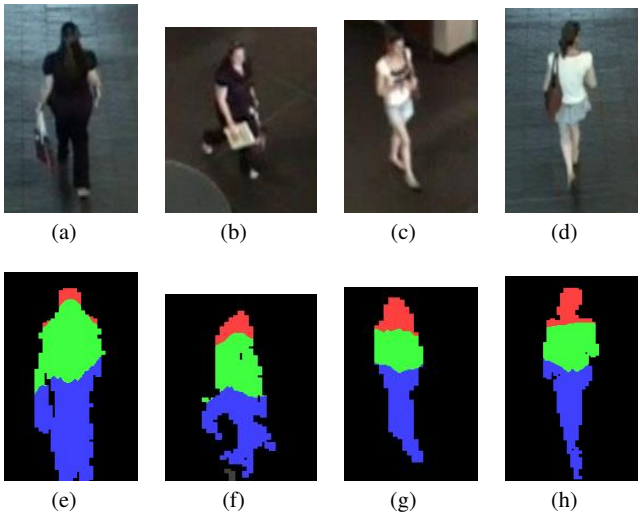


Fig. 3 Segmenting a person into head, torso and leg regions. The top row shows input colour images, the bottom row shows the segmented silhouettes. Red regions are designated as being the head, green as the torso and blue as the legs.

where i is the input value, B_w is the width of a single bin, b_n is the centre of bin n, $d_{i,n}$ is the distance between the input value i and the centre of bin n, and $w_{i,n}$ is the weight that is added to bin n in relation to sample i. M is the total number of bins that the sample is being split between, and m is an index into this set. This process of assigning weights is also illustrated in Figure 4.

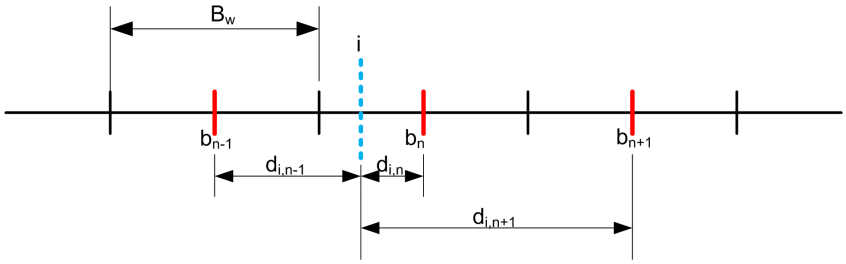


Fig. 4 Assigning weights to neighbouring bins in a soft histogram - The sample, i (blue dotted line), lies within bin n. Weight is assigned to bins n, $n-1$ and $n+1$ based on its proximity to the centre of each bin (red line). It can be seen that the maximum distance possible between a sample and the centre of an adjacent bin is $\frac{3}{2}B_w$.

A moving average of the histogram is calculated such that,

$$C'(t) = \frac{L-1}{L} \times C'(t-1) + \frac{C(t)}{L}, \tag{5}$$

where $C'(t)$ is the value of the average histogram at time t, $C(t)$ is the histogram computed for the frame at time t, and L is the learning rate. L is set according to

$$L = \frac{1}{T}; for\ T < W, \tag{6}$$

$$L = \frac{1}{W}; for\ W >= T, \tag{7}$$

where W is the number of frames used in the model, and T is the number of updates performed on the model. This ensures that the image that is used to initialise the model does not dominate the model for a significant number of frames. Instead, new information is incorporated quickly when the model is new to provide a better representation of the object being modelled.

Histograms are compared using the Bhattacharyya coefficient,

$$B(C^i, C^j) = \sum_{1}^{N} \sqrt{C^i(n) \times C^j(n)}, \tag{8}$$

where $B(C^i, C^j)$ is the Bhattacharyya coefficient for the comparison of the histograms C^i and C^j, $C^i(n)$ is the nth bin for the histogram C^i, and N is the total number of bins in the histogram. The histogram comparison is performed using histograms with their bin weights normalised such that they sum to 1. This is done to ensure size invariance. The comparison will return 1 for a perfect match, and 0 for no match.

When comparing colour models for two people, the similarity score is taken as the average of the three histogram comparisons,

$$P_{colour}(i,j) = \frac{(B(C^i_{Head}, C^j_{Head}) + B(C^i_{Torso}, C^j_{Torso}) + B(C^i_{Legs}, C^j_{Legs}))}{3}, \tag{9}$$

where $P_{colour}(i,j)$ is the similarity score between models i and j.

3.2 Height

The height of the person is used as a simple descriptor. The height is view invariant, whilst other dimensions (width and depth/thickness) are dependent on the camera angle as well as the persons pose (i.e. as a person walks their width changes as their legs move). Heights are stored for the head, torso and legs.

To determine the height of the person, the head and feet must be located in the image. Top and bottom contours of the motion image are extracted and smoothed. The top of the head is located by searching the top contour to determine the highest

point, x_h, y_h. The feet positions, $x_{lf}, y_{lf}, x_{rf}, y_{rf}$, are determined by finding the lowest point on the bottom contour either side of the head position. A mean foot position (x_f, y_f) is computed from these two detected feet locations.

The subject is separated into head, torso and legs as outlined in Section 3.1. The mean points along the two dividing contours,

$$x_{neck}, y_{neck} = \frac{\sum V_{neck}(x)}{N_{neck}}, \frac{\sum V_{neck}(y)}{N_{neck}}, \tag{10}$$

$$x_{waist}, y_{waist} = \frac{\sum V_{waist}(x)}{N_{waist}}, \frac{\sum V_{waist}(y)}{N_{waist}}, \tag{11}$$

$$\tag{12}$$

where x_{neck}, y_{neck} and x_{waist}, y_{waist} are the mean points of the neck and waist contours; and V_{neck} and V_{waist} are the contours that define the neck and waist boundaries (see Equation 2); are located and used to divide the region for the height calculations.

Figure 5 shows an example of the located head and feet points, and the points used to divide the subject into head, torso and legs.

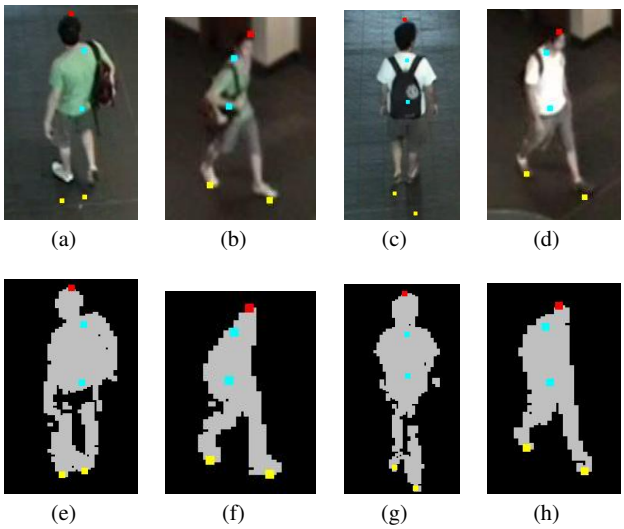


Fig. 5 Detecting the head and feet, and dividing the subject into three regions - The top row shows the colour input image and the bottom row shows the corresponding silhouette. Head, feet, waist and neck points are overlayed on both images. The head points are shown in red, feet shown in yellow, and median position of the waist and neck divisions shown in cyan. It can be seen that in some instances, shadows result in the feet being incorrectly located.

Using camera calibration, the image coordinates of the points can be transferred into a real world coordinate scheme. x_f, y_f is projected at an assumed z coordinate (height about the ground plane) of 0. The real world location of the feet can then be

projected back into the image plane at various heights (values of z) to determine the height above the ground plane of the waist, neck and head. Heights for the individual components can then be determined,

$$H_{head} = z^w_{head} - z^w_{neck}, \tag{13}$$
$$H_{torso} = z^w_{neck} - z^w_{waist}, \tag{14}$$
$$H_{legs} = z^w_{waist} - z^w_{feet}, \tag{15}$$

where H_{head}, H_{torso} and H_{legs} are the head, torso and legs heights in world coordinates, and z^w_{head}, z^w_{neck}, z^w_{waist}, z^w_{feet} are the world coordinates (height off the ground plane) of the head, neck, waist and feet. z^w_{feet} is always set to 0 (i.e. the person's feet are on the ground).

Heights are progressively updated over multiple observations,

$$H'_{head}(t) = \frac{L-1}{L} \times H'_{head}(t-1) + \frac{H_{head}(t)}{L}, \tag{16}$$
$$H'_{torso}(t) = \frac{L-1}{L} \times H'_{torso}(t-1) + \frac{H_{torso}(t)}{L}, \tag{17}$$
$$H'_{legs}(t) = \frac{L-1}{L} \times H'_{legs}(t-1) + \frac{H_{legs}(t)}{L}, \tag{18}$$

where $H'_{head}(t)$, $H'_{torso}(t)$ and $H'_{legs}(t)$ are the average head, torso and leg heights for the model at time t; $H_{head}(t)$, $H_{torso}(t)$ and $H_{legs}(t)$ are the heights for the image at the current time step computed as described in Equations 13 to 15, and L is the learning rate. L is defined in the same manner as the colour model (see Equations 6 and 7).

For each update, an error is calculated between the average of the model and the new observation,

$$F^e_{head}(t) = \left|H'_{head}(t) - H_{head}(t)\right|, \tag{19}$$
$$F^e_{torso}(t) = \left|H'_{torso}(t) - H_{torso}(t)\right|, \tag{20}$$
$$F^e_{legs}(t) = \left|H'_{legs}(t) - H_{legs}(t)\right|, \tag{21}$$

where $F^e_{head}(t)$, $F^e_{torso}(t)$ and $F^e_{legs}(t)$ are the frame errors for the head, torso and leg heights. Over time, an average error for each component (E_{head}, E_{torso} and E_{legs}, for the head, torso and legs respectively) can be computed from these frame errors ($F^e_{head}(t)$, $F^e_{torso}(t)$ and $F^e_{legs}(t)$) using the update method of the colour and height models (see Equations 5, 6 and 7). The cumulative error is used as an approximation to the standard deviation (it is assumed that the observations over time form a Gaussian distribution) of the error, as it is not practical to re-compute the standard deviation each frame, and not ideal to assume a fixed standard deviation. Given that the standard deviation for a sample set is defined as,

$$\sigma = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\mu - s_n)^2}, \tag{22}$$

and in the proposed model, for each measure there is one observation at each time step ($N = 1$), so the standard deviation at a given time step is,

$$\sigma = \sqrt{(\mu - s)^2} = \left|H'(t) - H(t)\right|, \tag{23}$$

which is the proposed error measure.

When comparing two size models, the mean heights and approximated standard deviations are used to determine the probability of a match. The probability for head, torso and legs heights are defined as,

$$P_{head}(i,j) = \Phi_{0,E_{head}(i)}(\left|H'_{head}(i) - H'_{head}(j)\right|), \tag{24}$$

$$P_{torso}(i,j) = \Phi_{0,E_{torso}(i)}(\left|H'_{torso}(i) - H'_{torso}(j)\right|), \tag{25}$$

$$P_{legs}(i,j) = \Phi_{0,E_{legs}(i)}(\left|H'_{legs}(i) - H'_{legs}(j)\right|), \tag{26}$$

where $P_{head}(i,j)$ is match between the head component two models, i and j, $E_{head}(i)$ is the approximated standard deviation of the head height for model i, $H'_{head}(i)$ is the mean head height for model i, and$\Phi_{\mu,\sigma}$ is the cumulative density function for the Gaussian distribution. The average of these scores,

$$P_{height}(i,j) = \frac{P_{head}(i,j) + P_{torso}(i,j) + P_{legs}(i,j)}{3}, \tag{27}$$

is taken as the match between models i and j.

3.3 Luggage

Luggage can be detected by observing symmetry in a person's silhouette (see [29]). From all angles, a person's silhouette is approximately symmetrical when they are walking or running, however objects they may be carrying lead to regions of asymmetry.

The head is located in the same manner as described in Section 3.2, and is used to divide the subject into two non-equal halves. Any region directly either side of the silhouette that has a height with 20% of the head height is not considered when calculating symmetry. We reason that (assuming the head is correctly detected) this portion of the silhouette will always contain the head and torso and some amount of the legs (depending on pose). Any luggage that is being carried will not clearly be visible, any asymmetry that is detected in this region is far more likely to arise from errors in the motion segmentation. Thus, this region is ignored.

This leaves us with two image regions, one either side of the centre column about the head, which are compared for symmetry. We define these image regions as being all pixels in the range $1 : Left, 1 : Y$ for the left portion, and $Right : X, 1 : Y$ for the right portion, where X and Y are the horizontal and vertical dimensions, and $Left$ and $Right$ are the left and right edges of the centre region. Asymmetri-

cal motion regions are located by searching for differences between the left portion and the mirror image of the right portion. Algorithmically, this is achieved as follows:

for $i = Left; i >= 1; i = i - 1$ **do**
  for $j = 1; j <= Y; j = j + 1$ **do**
    if $(M(i,j) = M(Right + (Left - i), j)$ **then**
      $A(i,j) = A(Right + (Left - i)) = False$
    else
      $A(i,j) = M(i,j)$
      $A(Right + (Left - i)) = M(i,j)$
    end if
  end for
end for

where A is a boolean image indicating regions of asymmetry and M is the motion image. It should be noted that the left and right portions that we are comparing are not necessarily the same size, and additional bounds checks that are not shown above need to be performed.

Connected components within A are detected, and the largest component is selected. If the size of this object is above a threshold (greater than 5% of the pixel count of the entire silhouette), it is accepted as a valid item of luggage. Examples of detected luggage items are shown in Figure 6. While the proposed approach is effective when the motion mask used is free of errors, it can be seen that when segmentation errors such as shadows are present (image pair (f)-(m)), additional regions of asymmetry result which can be incorrectly classified as an item of luggage. We also observe, that when a subject carries luggage on both sides of the body (such as image pairs (a)-(g) and (b)-(h)), the detector will not consistently detect the same item of luggage.

For the detected luggage items, we seek to model the colour of the item and the location of the luggage relative to the person. Two histograms are computed that correspond to the portion of the luggage item above the waist of the person, C_a, and the portion below the waist, C_b. The average amount of the luggage in each half is also stored (w_a and w_b). We do not consider whether the luggage is carried on the left or right side of the person as this is dependant on the angle of camera used. Soft histograms as used in the colour model (see Section 3.1) are used, and the model is updated in the same manner (see Equations 5, 6 and 7).

Luggage models are compared by performing a weighted sum of the histogram matches, with weights determined according to the amount of the luggage that exists in each region,

$$P_{luggage}(i,j) = min(w_a^i, w_a^j)B(C_a^i, C_a^j) + min(w_b^i, w_b^j)B(C_b^i, C_b^j), \tag{28}$$

where i and j are the two luggage models we are comparing, and *min* is an operator that determines the lowest of two values. In the event that one of models i and j has not detected a luggage item (and thus does not have a luggage model), $P_{luggage}(i,j) = 1$.

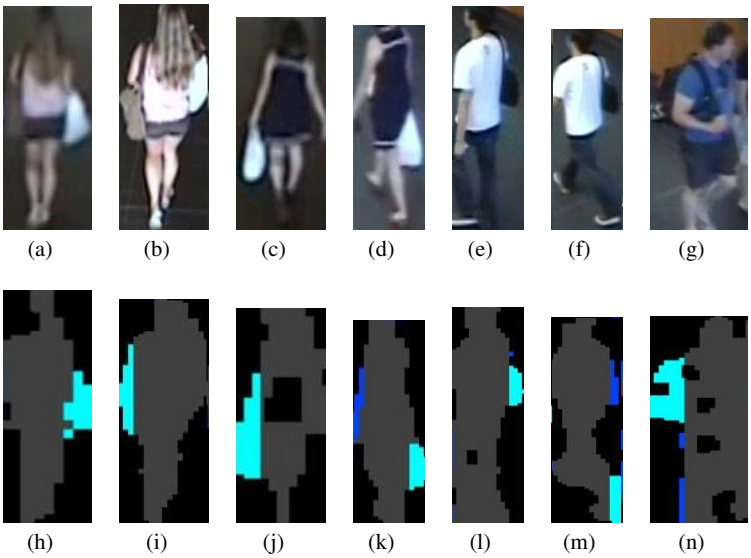


Fig. 6 Detecting luggage - Top row shows input colour image, bottom row shows silhouette images with detected luggage objects marked in Cyan. Areas of asymmetry that are not detected as luggage objects are shown in blue. Errors in detection tend to occur when a subject is carrying multiple items (such as (a)-(g) and (b)-(h)), or when segmentation errors are present (see (f)-(m)).

4 Soft Biometrics for Operational Analytics

The soft biometrics outlined in Section 3 can be used to match people across a disjoint camera network, however it is not practical or possible to match every person within the network. However as our aim is to determine operational statistics, we only require information from a representative subset of the population. As such, we propose a system that automatically selects a suitable subset of the population to match across a disjoint camera network.

The proposed soft biometrics require a series of observations to build accurate models. We use the object tracking system described in [17] to track objects within each camera view, and build soft biometric models (see Section 3). Objects are not tracked between views, and the relationship between cameras (i.e. distance between cameras) is unknown.

It should be noted that the performance of the soft biometrics is impacted by the performance of the object tracking system. Errors in the object tracking, such as incorrect segmentation will result in errors in the soft biometric models. In this situation, the part-based segmentation (see Section 3.1) may fail, or produce incorrect results. To help alleviate this, objects that are occluded (as determined by the object tracker) are not considered when building soft biometric models, however it

is expected that some erroneous models will still result due to errors in the object tracking.

The proposed system operates in two phases: a training phase and a testing phase. In the training phase, the soft biometrics that are built can be combined to construct a set of average soft biometrics. During operation, the proposed system then uses the average soft biometrics to automatically select subjects to attempt to match across the disjoint camera views. As the objects are matched, operational statistics, such as the time taken to traverse the environment, are gathered. The average biometrics are outlined in Section 4.1, and the proposed system for calculating operational statistics is presented in Section 4.2.

4.1 Average Soft Biometrics

We propose average soft biometrics for colour (see Section 4.1.1) and size (see Section 4.1.2). An average for luggage is not computed, as not all people carry luggage, and an item of luggage is not visible from all camera angles (i.e. a backpack can be seen from side on, but not from a front or rear view).

These average biometrics can then be used to determine how distinct a person is. When multiple soft biometrics are used, the most distinct mode is taken. We argue that as long as a person has one trait that is distinct, then that person is distinct (i.e. a person may be wearing clothes similar to many others, but if they are an unusual height, they are distinct).

4.1.1 Colour

The colour soft biometric consists of three colour histograms. The average biometric is simply the average histogram for each component,

$$A_c^{colour} = \frac{\sum_i^N C_{c,i}}{N}, \tag{29}$$

where A_c^{colour} is the average biometrics for component c, $C_{c,i}$ is the colour histogram of component c for subject i, and N is the number of subjects used to build the model.

The distinctiveness, D_i, of a given subject i, is determined in the same manner as two subjects are compared,

$$D_i = \frac{B(A_{head}^{colour}, C_{head,i}) + B(A_{torso}^{colour}, C_{torso,i}) + B(A_{legs}^{colour}, C_{legs,i})}{3}, \tag{30}$$

where $B(a,b)$ is the Bhattacharyya coefficient (see Equation 8) for the histograms a and b.

4.1.2 Size

Like the colour model, the size model consists of three parts: individual heights for the head, torso and legs. Each component is considered separately within the average model. Three histograms (A^{size}_{head}, A^{size}_{torso} and A^{size}_{legs}) are built that describe the likelihood that a given height for a given component will occur. Each histogram is normalised so that its bins sums to 1.

The distinctiveness of a subject can then be calculated as,

$$D_i = \frac{A^{size}_{head}(b_h) + A^{size}_{torso}(b_t) + A^{size}_{legs}(b_l)}{3}, \tag{31}$$

where b_h, b_t and b_l are the bins that correspond to the heights of the for the head, torso and legs in the subject's model.

4.2 Locating and Detecting Distinct People

The proposed system is intended for situations where there is a sparse disjoint camera network, and/or tracking throughout the complete environment is not possible. The tracking system outlined in [17] is used to track people within the individual cameras to build soft biometrics. An example of the tracking system output can be seen in Figure 7.

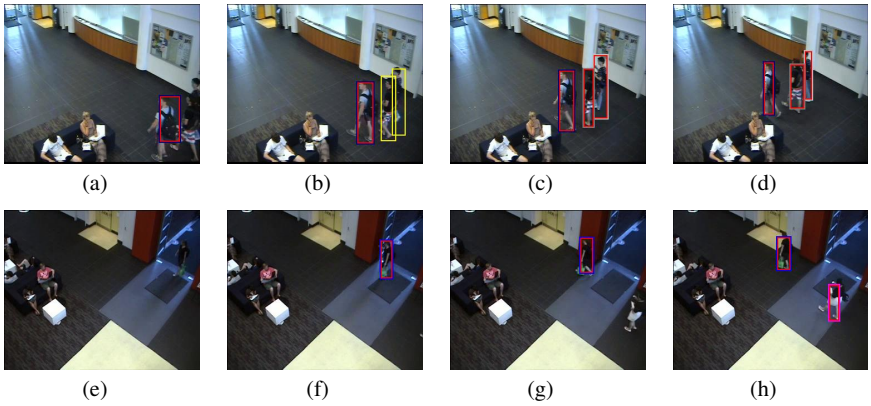
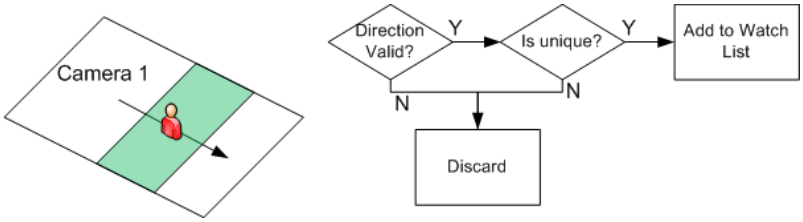


Fig. 7 Example tracking output - Top and bottom rows show two examples of output from the tracking system. Tracking output is not completely accurate, and errors in the tracking filter through to the construction of the soft biometrics.

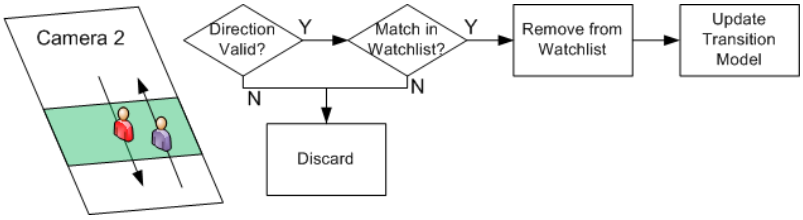
The proposed system operates in two stages:

1. Detecting unique people;
2. Detecting matches for the unique people.

These two situations are shown in Figure 8.



(a) A person is added to the watch list



(b) A person is compared to the watch list

Fig. 8 Proposed system - The object tracking algorithm operates within the shaded regions of the camera views to build soft biometrics. People who are classified as unique are added to a watch list, which other people detected elsewhere in the network are compared to.

In the first stage, we seek to identify people who are distinct, and who can potentially be re-detected elsewhere in the camera network. Subjects are tracked as they move through the area of interest, and if they can be tracked for long enough to build a sufficient model (20 frames in the proposed system), they are compared to the average model to determine if they are unique. As outlined in Section 4.1, if any one mode is deemed to be unique, the person is considered unique. Unique people are added to a watch list so that they can be re-detected elsewhere. A minimum frame limit is imposed to reduce errors in the soft biometric models caused by tracking or segmentation errors, as we assume that the majority of frames will be correctly segmented.

In the second stage, we aim to match people to those previously detected. Like the first stage, people are tracked through the region of interest and once a model has been built, they are compared to all subjects in the watch list by comparing the soft biometric models. For two objects to be matched, the colour, height and luggage soft biometrics must all have a similarity above a threshold,

$$P_{colour}(i,j) \geq T_{colour} \;\&\; P_{height}(i,j) \geq T_{height} \;\&\; P_{luggage}(i,j) \geq T_{luggage}, \qquad (32)$$

where i and j and the source and target objects being matched; $P_{colour}(i,j)$, $P_{height}(i,j)$ and $P_{luggage}(i,j)$ are the matches between the two models as defined in Equations 9, 27 and 28 respectively; and T_{colour}, T_{height} and $T_{luggage}$ are thresholds for the three soft biometrics. We argue that as each of these soft biometrics is essentially a weak classifier, all three must be satisfied for a match to occur. Thresholds are selected according to the performance of each classifier, with T_{height} and $T_{luggage}$ typically set lower than T_{colour} (more false positives, fewer false negatives) as they are less reliable. In the event of multiple targets satisfying Equation 32, the target with the highest value for $P_{colour}(i,j)$ is selected as the match, as the colour modality is the most reliable.

The first valid match that an object receives is taken to be correct (i.e. the system does not wait to see if there is a better match later on). This can potentially lead to errors if people of a similar appearance are present at the same time. However the requirement for people to be 'distinct' should reduce the likelihood of such an occurrence. As matches are made, the transition information (mean and standard deviation of time taken) is calculated.

In both stages of the system, the direction the subject is moving is monitored to determine if the person is valid for a comparison, or to be added to the watch list. For example, whilst a person may be tracked through a region where unique people are being detected, if they are a moving in a direction that will exit the environment they will not be compared to the average model, or added to a watch list (see Figure 9).

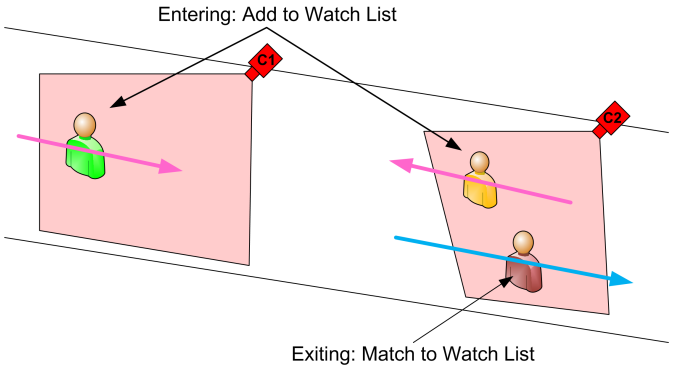


Fig. 9 Entering and exiting objects - In the scene shown either end of the corridor is an entrance and exit. Objects entering the scene are checked for uniqueness, and possibly added to the watch list. Objects exiting the scene are simply compared to the contents of the watch list.

When training an average biometric model, the system operates in the first mode only and disregards direction. Every object that can be tracked for a sufficient period to build a model is included in the average model.

5 Evaluation

We perform two evaluations:

1. Evaluate the performance of the soft biometrics themselves for a identification task;
2. Evaluate the performance of using soft biometrics to identify distinct looking people, and match these people across a disjoint camera network.

These evaluations are presented in Sections 5.1 and 5.2 respectively. Databases have been captured in house for each evaluation, and these are outlined in Sections 5.1.1 and 5.2.1 respectively.

5.1 Soft Biometric Performance

5.1.1 Evaluation Data

To evaluate the performance of the soft biometrics for recognising people in a surveillance environment, a database has been captured that contains people moving through a building environment by up to 8 surveillance cameras. In total, 80 people are present in the database.

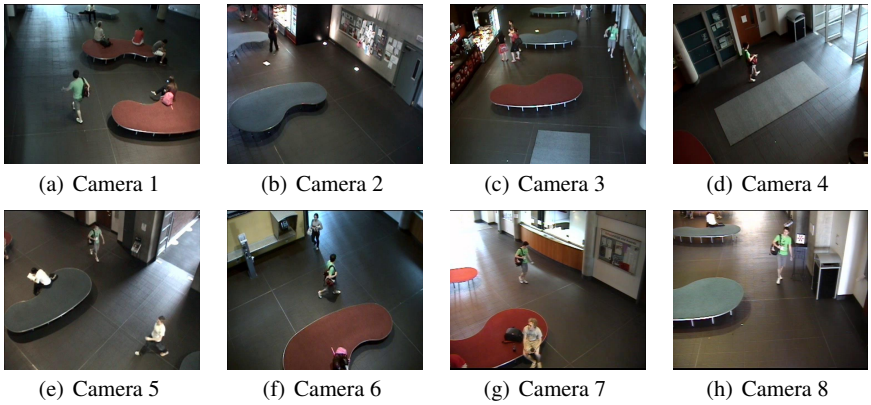
The vast majority of subjects will only pass through a subset of the camera network, and that subset varies from person to person. This provides a highly unconstrained environment in which to test the proposed soft biometrics. The cameras have all been calibrated using Tsai's method [61], and an example image from each camera is shown in Figure 10.

From Figure 10, it can be seen that there is varied lighting across the different camera views, and that subjects will be observed from different angles as they move through the network.

Subject locations in the network are hand annotated (every 10th frame is annotated and intermediate frame locations are interpolated). The annotated locations are coarse and the subject location is refined using motion segmentation (person detection [9, 63] could also be used), thus there are also potential errors that can be cause by incorrect motion segmentation.

In our evaluation, we consider the accuracy of the size and colour soft biometrics. We do not consider luggage as it is not carried by all subjects, and not visible from all camera angles. Cumulative match characteristic (CMC) and synthetic recognition rate (SRR) [24] curves are plotted for both the size and colour biometric as well as an equivalent colour biometric using a hard histogram, and for a fused system that uses a weighted sum to combine the colour (soft histogram), size and luggage biometrics. The following three configurations of soft biometric models are evaluated:

1. Soft biometric models trained on a single camera view;
2. Soft biometric models trained on two camera views;



(a) Camera 1 (b) Camera 2 (c) Camera 3 (d) Camera 4

(e) Camera 5 (f) Camera 6 (g) Camera 7 (h) Camera 8

Fig. 10 Example images from the 8 cameras used to capture the soft biometric evaluation database. Note that there are different lighting conditions in each camera, and that as person moves through the network they are viewed from several different angles.

3. Soft biometric models trained on three camera views.

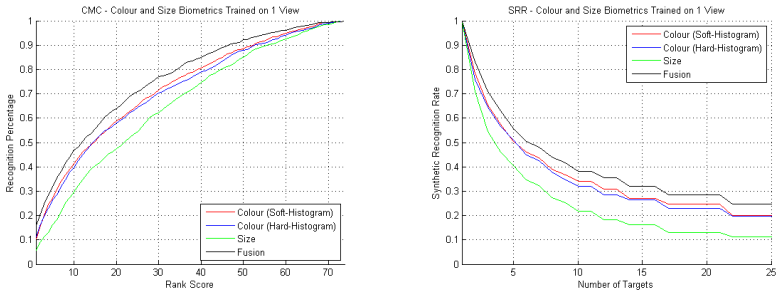
Models are trained and tested on separate views, and 20 images are used from each camera view. It should be noted that as we increase the number of views required for training models, fewer subjects from the database are included as subjects with insufficient data are omitted. As subjects within the database take widely varying paths through the camera network and pass through different numbers of cameras, not all subjects can be used in all cases.

5.1.2 Results

Cumulative match characteristic (CMC) and synthetic recognition rate (SRR) plots are shown for the colour soft biometric (hard and soft histogram variants), size biometric, and a fused system trained on one, two and three camera views are shown in Figures 11, 12 and 13 respectively.

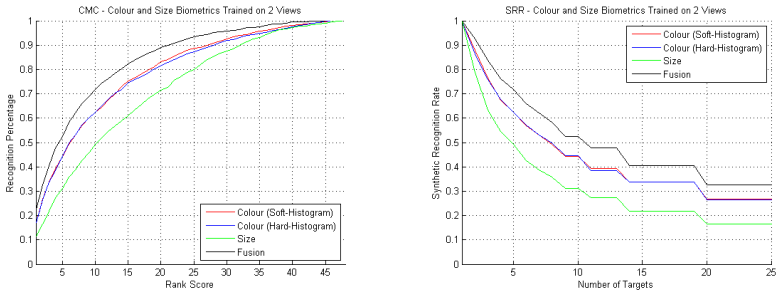
It can be seen that the colour model consistently outperforms the height model, and both models improve as more views are used to train the models. The soft histogram variant of the colour model offers a small but consistent performance advantage over the hard histogram equivalent.

The superior performance of the colour model is to be expected, as there is a far greater variation in colour than size. A large number of people have heights that are within a few centimetres of each other, meaning there is much less variation in this trait. The height biometric is also less accurate to extract. Segmentation errors in the silhouette will result in errors in the detected height. An error of only a few pixels can result in a difference of a few centimetres or more, depending on where in the image the subject appears. While the colour biometric is also susceptible



(a) CMC (b) SRR

Fig. 11 Cumulative match characteristic (CMC) and synthetic recognition rate (SRR) plots for models trained on a single camera view.

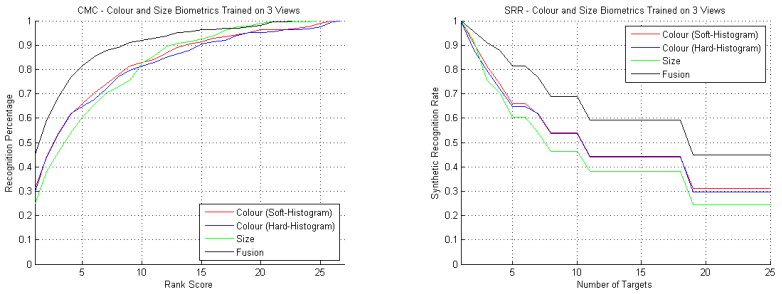


(a) CMC (b) SRR

Fig. 12 Cumulative match characteristic (CMC) and synthetic recognition rate (SRR) plots for models trained on two camera views.

to segmentation errors, these do not result in a significantly different observation unless the segmentation errors result large portions of the person not being visible (i.e. their legs or torso are not detected), or a large portion of the background being included in the model. In cases such as these, it is likely that the division of the person into head, torso and legs would fail (see Section 3.1) and a model would not be built for that frame.

It can also be seen that the soft histogram colour model consistently performs as well as, or better than, the hard histogram equivalent. While using a soft histogram only offers a small performance improvement, this is not surprising as the underlying model based on the division of the subject into three parts is unchanged and the scope for improvement using the soft histogram is limited. The nature of the soft histogram though, which allows ambiguous samples to be spread across multiple bins, does offer some help when dealing with challenging lighting conditions. Colours which lie close to a bin boundary are distributed across neighbouring bins, providing a limited amount of invariance to changes in colour balance between cam-



(a) CMC (b) SRR

Fig. 13 Cumulative match characteristic (CMC) and synthetic recognition rate (SRR) plots for models trained on three camera views.

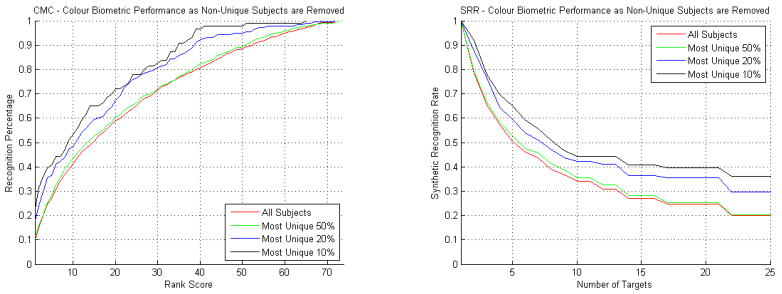
eras. However, if the change in a colour between cameras is too large (greater than 1.5 bin widths in the proposed system, see Section 3.1), the soft histogram will offer no support in dealing with colour changes.

We also observe that performance of the soft biometrics improve are more views are used to build the models. This is to be expected, as the use of multiple views ultimately leads to a more complete and accurate model. In the case of the colour biometric, a person may appear different from the front or back due to an item they are carrying, or an item of clothing such as a jacket that is open at the front. Including data from multiple viewpoints allows for a model that better represents the persons overall appearance.

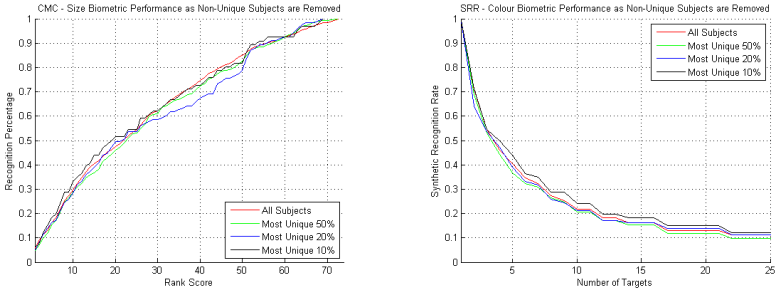
From Figures 11, 12 and 13, it can also be seen that a simple weighted summation of the modalities results in a noticeable improvement in performance.

Figure 14 shows CMC and SRR plots for both the colour and size soft biometrics as non-unique subjects are removed from the probe set (note that the non-unique subjects are still contained in the gallery set). Subjects who have a uniqueness (determined using an average biometric, see Section 4.1 for further details) below a threshold are removed from the probe set. In Figure 14, soft biometric models are trained and tested on data captured from a single camera view.

For the colour modality, it can be seen that as the less distinct subjects are removed from the probe set, the matching accuracy improves and the performance gain becomes more pronounced as an increasing number of probe samples are excluded. The size modality however does not offer the same improvement, and at some operating points performance of the more distinct probe set is less than that of the unfiltered set. We attribute this to the inherent inaccuracy in the size models, and the low variation in peoples heights. As non-distinct subjects are removed, some of the more ‘unique’ subjects that remain are likely to have heights that are the result or poor segmentation or other errors, thus making them more difficult, rather than easier, to match correctly. However, we do note that at low ranks a performance improvement (while not as large as that in the colour domain) is observed. As our



(a) CMC, Colour (b) SRR, Colour



(c) CMC, Size (d) SRR, Size

Fig. 14 CMC and SRR plots for the colour and size soft biometrics as non-unique subjects are removed from the gallery set.

target application is intended to operate with smaller number of subjects (i.e. 20 or less) in the gallery at any given time, we argue there is still value in using the height modality.

5.2 Soft Biometrics for Operations Tasks

5.2.1 Evaluation Data

A small database is captured in house, consisting of data captured from up to four cameras. Figure 15 shows the approximate layout of the camera network. All cameras are recorded simultaneously at 25 frames per second.

Two data sets are captured:

- Data set 1 uses only cameras 1 and 2, and consists of three 15,000 frame (10 minute, 25fps) sequences captured at different times (0915, 1315, 1615) during a single day. The three data sets have increasing levels of pedestrian traffic (0915 has the fewest people, 1615 the most people). These sequences are stored and

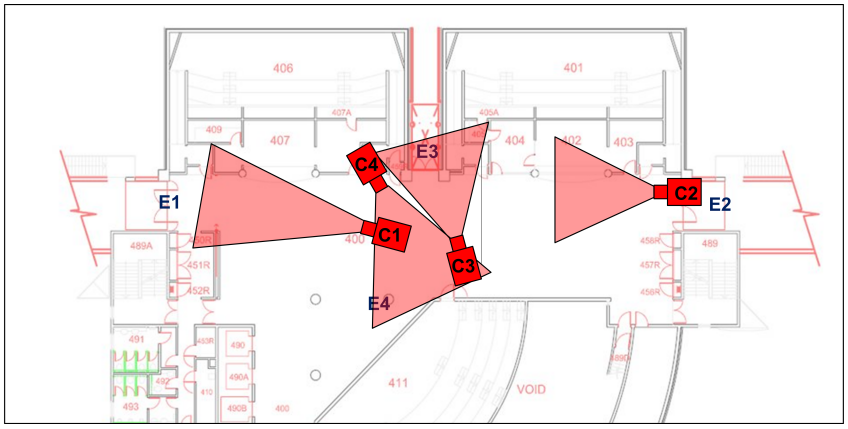


Fig. 15 Camera network layout - Four cameras (C1, C2 ,C3 and C4; approximate camera views shown in red) observe the three external doorways(E1, E2 and E3), and an internal door. People enter and exit through these doors, as well as through other internal doorways and spaces within the building.

processed at 4CIF resolution (704x576). These three sequences will be referred to as DS1-06-0915, DS1-06-1315 and DS1-06-1615.

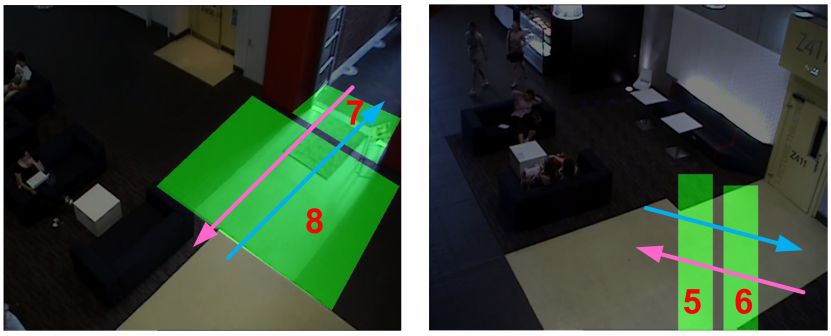
- Data set 2 uses all four cameras, and has two sequences of 30,000 frames each. These sequences have been captured at the same time of day (1710) on consecutive days. Each sequences contains a similar amount of traffic (roughly equivalent to DS1-06-1615). These sequences are stored and processed at CIF resolution (352x288). These two sequences will be referred to as DS2-17-1710 and DS2-18-1710.

Each camera view is configured with entrance and exit directions, as outlined in Section 4.2. Entrances and exits are shown in Figure 5.2.1. It should be noted that people who enter through one camera can exit the scene through any other camera, or through an area that is not covered by the cameras (see Figure 15). In this evaluation, we restrict each camera to having a single entrance and exit. We do this to increase the number of samples we observe for each transition, which allows us to better evaluate the accuracy of the estimation. It should be noted though that the proposed framework can support multiple entrances and exits within each camera, and it would also be possible to automatically detect these (see [60]), however this is beyond the scope of this work.

In our evaluation, average biometric models are trained on a single view at a given time period, and the system is then tested using data using captured at the other time period(s). Camera 1 is always used as the training view as it receives the most pedestrian traffic, and so provides the most training data and best average models.

People are detected entering and existing all views within the test set (see Figure 5.2.1), resulting in estimates for every possible path through the camera network. In

(a) Camera 1 (b) Camera 2



(c) Camera 3 (d) Camera 4

Fig. 16 Camera configuration - Entrances and exits over which people are monitored are shown. Purple arrows indicate an entrance, blue arrows indicate an exit.

our evaluation we consider the accuracy of these estimates, and the accuracy with which matches are made between the camera views.

5.2.2 Results

Evaluation results for the proposed system using data set 1 are shown in Table 1.

It can be seen from Table 1 that the proposed approach is able to accurately estimate the travel time in both directions in the majority of cases, particularly when sufficient samples are observed. Due to the fact that DS1-06-0915 contains the fewest people, estimates based on this data set are the least accurate. Similarly, estimates obtained using DS1-06-1615 are closest to the ground truth due to the increased number of people to detect and match.

Examples of correct and incorrect matches made using data set 1 are shown in Figures 17 and 18 respectively. It can be seen that the proposed soft biometrics are able to cope with variation in the viewing angle, as well as some variation in lighting

Fig. 17 Examples of correctly matched people, all for $T_m = 0.6$. Top row shows an example of the 'source' image, bottom row the 'target' image. Image pairs (a)-(g) and (b)-(h) are from DS1-06-0915, (c)-(i) and (d)-(j) from DS1-06-1315, and (e)-(k) and (f)-(l) from DS1-06-1615. People are correctly matched despite changes in pose, and changes in colour caused by the differing lighting conditions.

Fig. 18 Examples of incorrectly matched people, all for $T_m = 0.6$. Top row shows an example of the 'source' image, bottom row the 'target'. Image pairs (a)-(g) and (b)-(h) are from DS1-06-0915, (c)-(i) and (d)-(j) from DS1-06-1315, and (e)-(k) and (f)-(l) from DS1-06-1615. Errors are made either due to poor image segmentation ((a)-(g), (c)-(i), (d)-(j)), or due to the people having a similar appearance ((b)-(h), (d)-(j), (e)-(k), (f)-(l)).

Distinct Threshold	Training Data set	Testing Data set	C1⇒C2		C2⇒C1	
			Time Taken (frames)	Match Rate	Time Taken (frames)	Match Rate
$T_m = 0.5$	0915	1315	$\mu = 471, \sigma = 216$	60% (6/10)	$\mu = 431, \sigma = 135$	75% (9/12)
	0915	1615	$\mu = 342, \sigma = 112$	86% (6/7)	$\mu = 410, \sigma = 146$	57% (12/21)
	1315	0915	$\mu = 269, \sigma = 167$	33% (1/3)	$\mu = 342, \sigma = 137$	80% (4/5)
	1315	1615	$\mu = 342, \sigma = 112$	86% (6/7)	$\mu = 410, \sigma = 146$	57% (12/21)
	1615	0915	$\mu = 269, \sigma = 167$	33% (1/3)	$\mu = 342, \sigma = 137$	80% (4/5)
	1615	1315	$\mu = 486, \sigma = 239$	63% (5/8)	$\mu = 388, \sigma = 41$	75% (9/12)
$T_m = 0.6$	0915	1315	$\mu = 601, \sigma = 271$	66% (2/3)	$\mu = 395, \sigma = 44$	73% (8/11)
	0915	1615	$\mu = 340, \sigma = 121$	83% (5/6)	$\mu = 365, \sigma = 122$	61% (11/18)
	1315	0915	$\mu = 269, \sigma = 167$	33% (1/3)	$\mu = 338, \sigma = 153$	75% (3/4)
	1315	1615	$\mu = 340, \sigma = 121$	83% (5/6)	$\mu = 404, \sigma = 153$	61% (11/18)
	1615	0915	$\mu = 269, \sigma = 167$	33% (1/3)	$\mu = 338, \sigma = 153$	75% (3/4)
	1615	1315	$\mu = 580, \sigma = 237$	75% (3/4)	$\mu = 379, \sigma = 62$	73% (8/11)
$T_m = 0.7$	0915	1315	$\mu = 674, \sigma = 307$	50% (1/2)	$\mu = 387, \sigma = 47$	50% (2/4)
	0915	1615	$\mu = 420, \sigma = 48$	100% (2/2)	$\mu = 420, \sigma = 118$	78% (7/9)
	1315	0915	$\mu = 106, \sigma = 0$	0% (0/1)	$\mu = 421, \sigma = 247$	33% (1/3)
	1315	1615	$\mu = 399, \sigma = 49$	100% (3/3)	$\mu = 430, \sigma = 170$	69% (9/13)
	1615	0915	$\mu = 106, \sigma = 0$	0% (0/1)	$\mu = 587, \sigma = 97$	50% (1/2)
	1615	1315	$\mu = 674, \sigma = 307$	50% (1/2)	$\mu = 377, \sigma = 46$	60% (3/5)

Table 1 Evaluation results for data set 1 (DS1-06-0915 is abbreviated to 0915, DS1-06-1315 to 1315 and DS1-06-1615 to 1615) - Three different thresholds for selecting distinct people are used. Lower thresholds result in a higher number of people being added to the watch list and matched, however there are more incorrect matches made between people with a similar appearance. Ground truth times (measured in frames) taken to move through the network are $\mu = 419, \sigma = 62$ for Left to Right, and $\mu = 407, \sigma = 51$ for Right to Left.

conditions. The incorrect matches (see Figure 18) are caused by either the source and target having a similar appearance; or by poor segmentation, either from the object tracking or motion segmentation (or both).

Results for data set 2 are shown in Table 2. For simplicity, we consider transitions between the same cameras but in different directions as the same (i.e. the transitions C1⇒C2 and C2⇒C1 are considered together as C1⇔C2).

Examples of correct and incorrect matches obtained using data set 2 are shown in Figures 19 and 20 respectively. As was observed in data set 1, the soft biometrics are able to cope with variations in viewing angle and colour balance between the camera views. This is even more apparent in data set 2 as cameras 3 and 4 have a greater variation in viewing angle relative to cameras 1 and 2, then cameras 1 and 2 do to each other. Within data set 2, this problem is further exacerbated by the reduction in resolution from 4CIF to CIF.

As was observed in data set 1, a significant portion of the errors are due to the tracking errors (see Figure 7 for an example of the tracking errors present) resulting in soft biometric models that are actually a combination of one or more people. Examples of this can be seen in Figure 20 (see image pairs (e)-(l), (f)-(m) and (g)-(n)). Despite these incorrect matches, accurate estimates of travel times between the camera views can still be obtained provided a sufficient number of samples

Distinct Threshold	Transition	Training D2-14-1710 Testing D2-15-1710		Training D2-15-1710 Testing D2-14-1710	
		Time Taken (frames)	Match Rate	Time Taken (frames)	Match Rate
$T_m = 0.6$	C1⇔C2	$\mu = 439, \sigma = 181$	73% (30/41)	$\mu = 459, \sigma = 153$	74% (23/31)
	C1⇔C3	$\mu = 451, \sigma = 235$	25% (1/4)	$\mu = 369, \sigma = 226$	50% (1/2)
	C1⇔C4	$\mu = 849, \sigma = 0$	0% (0/1)	$\mu = 974, \sigma = 0$	0% (0/1)
	C2⇔C3	$\mu = 215, \sigma = 88$	67% (2/3)	$\mu = 302, \sigma = 168$	50% (2/4)
	C2⇔C4	$\mu = 376, \sigma = 52$	67% (2/3)	$\mu = 278, \sigma = 52$	50% (3/5)
	C3⇔C4	N/A	N/A (0/0)	N/A	N/A (0/0)
$T_m = 0.7$	C1⇔C2	$\mu = 423, \sigma = 179$	68% (26/38)	$\mu = 459, \sigma = 153$	76% (22/29)
	C1⇔C3	$\mu = 451, \sigma = 235$	25% (1/4)	$\mu = 369, \sigma = 226$	50% (1/2)
	C1⇔C4	$\mu = 849, \sigma = 0$	0% (0/1)	$\mu = 974, \sigma = 0$	0% (0/1)
	C2⇔C3	$\mu = 215, \sigma = 88$	67% (2/3)	$\mu = 302, \sigma = 168$	50% (2/4)
	C2⇔C4	$\mu = 376, \sigma = 52$	67% (2/3)	$\mu = 278, \sigma = 52$	50% (3/5)
	C3⇔C4	N/A	N/A (0/0)	N/A	N/A (0/0)
$T_m = 0.8$	C1⇔C2	$\mu = 436, \sigma = 121$	75% (6/8)	$\mu = 555, \sigma = 306$	33% (2/6)
	C1⇔C3	$\mu = 441, \sigma = 101$	0% (0/2)	$\mu = 369, \sigma = 226$	50% (1/2)
	C1⇔C4	N/A	N/A (0/0)	$\mu = 974, \sigma = 0$	0% (0/1)
	C2⇔C3	N/A	N/A (0/0)	$\mu = 302, \sigma = 168$	50% (2/4)
	C2⇔C4	N/A	N/A (0/0)	$\mu = 487, \sigma = 220$	33% (1/3)
	C3⇔C4	N/A	N/A (0/0)	N/A	N/A (0/0)

Table 2 Evaluation results for data set 2 - Three different thresholds for selecting distinct people are used. Like data set 1, lower thresholds result in a higher number of people being added to the watch list and matched. Ground truth transition times are $\mu = 413, \sigma = 57$ for C1⇔C2, $\mu = 236, \sigma = 233$ for C1⇔C3, $\mu = 242, \sigma = 47$ for C1⇔C4, $\mu = 259, \sigma = 25$ for C2⇔C3 and $\mu = 290, \sigma = 65$ for C2⇔C4. No ground truth transition time for C3⇔C4 is calculated as insufficient examples are observed in the data (there is only 1 occurrence).

can be observed. For transitions such as camera 1 to camera 2 (which is the main thoroughfare in this environment) where there is a large number of observations, the overall estimate of the travel time is still accurate despite the errors. For the other less common transitions however, estimates are much less accurate. However if additional training data was available it is reasonable to expect that these estimates would improve over time.

In both data sets, we observe that as expected, increasing T_m reduces the number of people matched between views. However the overall accuracy of the matches remains consistent in most cases. In data set 1 the overall match rate is 66% for $T_m = 0.5$, 67% for $T_m = 0.6$ and 64% for $T_m = 0.7$. In data set 2 we observe that the match rate is 62% for $T_m = 0.6$, 66% for $T_m = 0.7$ and 46% for $T_m = 0.8$.

As T_m increases, the number of erroneous matches made due to people having a non-distinct appearance decreases, however errors caused by poor tracking and segmentation persist. This is particularly apparent in data set 2 when $T_m = 0.8$. Figure 21 shows 7 of the 10 errors obtained when training using DS2-15-1710 and testing using DS2-14-1710. It can be seen that in all errors the source is poorly segmented. Either part of the subject has been clipped ((a), (c), (f), (g)), or the subject is actually two people ((b), (d), (e)). Poorly or erroneously segmented people are more readily retained as T_m increases as they simply appear more unique. The poor segmentation

Fig. 19 Examples of correctly matched people, all for $T_m = 0.6$. Top row shows an example of the 'source' image, bottom row the 'target'. Image pairs (a)-(g), (b)-(h) and (c)-(i) show correct matches made between cameras 1 and 3; pairs (d)-(j) and (e)-(k) show correct matches made between cameras 2 and 3; and pair (f)-(l) shows a match made between cameras 2 and 4.

Fig. 20 Examples of incorrectly matched people, all for $T_m = 0.6$. Top row shows an example of the 'source' image, bottom row the 'target' image. Image pairs (a)-(h), (b)-(i), (c)-(j) and (d)-(k) are matched incorrectly due to the source and target being of a similar colour. The erroneous image pairs (e)-(l), (f)-(m) and (g)-(n) are the result of poor tracking leading to poorly trained soft biometric models.

of these subject means that the soft biometrics are less accurate and so the subject themselves are more likely to be erroneously matched. This could be overcome either through more accurate tracking that feeds into the soft biometrics (which could possibly be aided by using a technique such as the 'virtual gate' [35] to detect people as they cross the region of interest), or by applying more stringent checks to the segmented people to make sure that they do include only a single, complete person.

Fig. 21 Incorrect matches when training with D2-15-1710 and testing with D2-14-1710 for $T_m = 0.8$. The top row shows an example source image for the subject. The bottom row shows an example target image for the subject.

6 Conclusion

In this chapter we have demonstrated how soft biometrics can be used to recognise people in a disjoint camera network, and how they can be applied to measure operational information such as the average time taken to travel between two points. We have presented a novel approach that calculates an average soft biometric that can be used to locate distinct or unusual looking people, allowing the system to automatically select an appropriate subset of people within a scene for measurement.

In this chapter we have applied this technique to estimating the travel time between multiple points within an environment. We have shown that using simple descriptors (colour, height and luggage), we are able to obtain good estimates for the transition times provided there are sufficient observations. By filtering the set of people that we seek to re-detect, we effectively simplify the matching problem, enabling these simple descriptors to be applied with good results. It should be noted

however that additional soft biometrics (such as hair and skin colour, gender, etc.) could also be used within this framework, as well as more complex texture based descriptors. The proposed framework also has additional applications, such as to estimating a coarse trajectory through an environment, or estimating dwell time within a space. The underlying soft biometrics also have many other applications within security tasks.

Based on this results contained within this chapter, there are several avenues for future work. These include:

- Extending the approach to an incremental framework, where the requirements on the 'uniqueness' of individuals is continuously relaxed as more observations are made, and the layout of the network becomes established.
- The use of a robust estimator to estimate the transitions times.
- Testing the proposed system on larger data sets and more complex camera networks.
- The inclusion of additional soft biometrics and descriptors, and improvements to the accuracy of the soft biometrics, through techniques such as colour normalisation to help overcome errors caused by inconsistent lighting within and between camera views.
- Investigating methods to ensure that only correctly segmented people are included, either through improvement of the underlying tracking system, through automatically detecting tracking errors, or through the use of other techniques to detect people as the enter the environment.

References

[1] Ailisto H, Vildjiounaite E, Lindholm M, Makela S, Peltola J (2006) Soft biometrics–combining body weight and fat measurements with fingerprint biometrics. Pattern Recognition Letters 27(5):325–334

[2] Bak S, Corvee E, Bremond F, Thonnat M (2010) Person re-identification using haar-based and dcd-based signature. In: 2nd Workshop on Activity Monitoring by Multi-Camera Surveillance Systems, AMMCSS 2010, in conjunction with 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS, AVSS

[3] Bak S, Corvee E, Bremond F, Thonnat M (2010) Person re-identification using spatial covariance regions of human body parts. In: Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on, pp 435–440

[4] Bazzani L, Cristani M, Perina A, Farenzena M, Murino V (2010) Multiple-shot person re-identification by hpe signature. In: Pattern Recognition (ICPR), 2010 20th International Conference on, pp 1413–1416

[5] Beymer D (2000) Person counting using stereo. In: Human Motion, Workshop on, IEEE Computer Society, Los Alamitos, CA, USA, vol 0, p 127, DOI http://doi.ieeecomputersociety.org/10.1109/HUMO.2000.897382

[6] Breitenstein MD, Reichlin F, Leibe B, Koller-Meier E, Gool LV (2010) Online multi-person tracking-by-detection from a single, uncalibrated camera. IEEE Transactions on Pattern Analysis and Machine Intelligence 99

[7] Chien SY, Chan WK, Cherng DC, Chang JY (2006) Human object tracking algorithm with human color structure descriptor for video surveillance systems. In: Multimedia and Expo, 2006 IEEE International Conference on, pp 2097–2100

[8] Collins R, Gross R, Shi J (2002) Silhouette-based human identification from body shape and gait. In: Proceedings of IEEE Conference on Face and Gesture Recognition, pp 351–356

[9] Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Schmid C, Soatto S, Tomasi C (eds) International Conference on Computer Vision & Pattern Recognition, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, vol 2, pp 886–893

[10] Damen D, Hogg D (2008) Detecting carried objects in short video sequences. In: Proceedings of the 10th European Conference on Computer Vision: Part III, Springer-Verlag, Berlin, Heidelberg, ECCV '08, pp 154–167, DOI http://dx.doi.org/10.1007/978-3-540-88690-7_12, URL http://dx.doi.org/10.1007/978-3-540-88690-7_12

[11] Damen D, Hogg D (2009) Recognizing linked events: Searching the space of feasible explanations. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp 927–934

[12] Dantcheva A, Velardo C, DAngelo A, Dugelay JL (2011) Bag of soft biometrics for person identification: New trends and challenges. Multimedia Tools and Applications 51(2):739–777

[13] Demirkus M, Garg K, Guler S (2010) Automated person categorization for video surveillance using soft biometrics. In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol 7667, p 22

[14] Demirkus M, Toews M, Clark JJ, Arbel T (2010) Gender classification from unconstrained video sequences. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, p 55–62

[15] Denman S, Fookes C, Bialkowski A, Sridharan S (2009) Soft-Biometrics: unconstrained authentication in a surveillance environment. Digital Image Computing: Techniques and Applications (DICTA) pp 196–203

[16] Denman S, Fookes C, Sridharan S (2009) Improved simultaneous computation of motion detection and optical flow for object tracking. In: Digital Image Computing: Techniques and Applications (DICTA), Melbourne, Australia

[17] Denman S, Fookes C, Sridharan S (2010) Group segmentation during object tracking using optical flow discontinuities. In: The 4th Pacific-Rim Symposium on Image and Video Technology, Singapore, pp 270–275

[18] Denman S, Fookes C, Sridharan S, Ryan D (2010) Multi-modal object tracking using dynamic performance metrics. In: 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Boston, USA, pp 286–293

[19] Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2010) Person re-identification by symmetry-driven accumulation of local features. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp 2360–2367

[20] Fleuret F, Berclaz J, Lengagne R, Fua PAFP (2008) Multicamera people tracking with a probabilistic occupancy map. Transactions on Pattern Analysis and Machine Intelligence 30(2):267–282, 0162-8828

[21] Forssen PE (2007) Maximally stable colour regions for recognition and matching. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pp 1 –8, DOI 10.1109/CVPR.2007.383120

[22] Gheissari N, Sebastian TB, Hartley R (2006) Person reidentification using spatiotemporal appearance. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol 2, pp 1528–1535

[23] Gray D, H T (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: ECCV '08: Proceedings of the 10th European Conference on Computer Vision, Marseille, France, pp 262–275

[24] Gray D, Brennan S, Tao H (2007) Evaluating appearance models for recognition, acquisition and tracking. In: PETS

[25] Hahnel M, Klunder D, Kraiss K (2004) Color and texture features for person recognition. In: IEEE International Joint Conference on Neural Networks, Budapest, Hungary, p 652
[26] Han J, Bhanu B (2006) Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence 28:316–322
[27] Haritaoglu I, Flickner M (2002) Attentive billboards: towards to video based customer behavior understanding. In: Applications of Computer Vision, 2002. (WACV 2002). Proceedings. Sixth IEEE Workshop on, pp 127 – 131, DOI 10.1109/ACV.2002.1182169
[28] Haritaoglu I, Cutler R, Harwood D, Davis LS (1999) Backpack: Detection of people carrying objects using silhouettes. Computer Vision, IEEE International Conference on 1:102, DOI http://doi.ieeecomputersociety.org/10.1109/ICCV.1999.791204
[29] Haritaoglu I, Harwood D, Davis L (2000) W4: real-time surveillance of people and their activities. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8):809 – 830
[30] Haritaoglu I, Beymer D, Flickner M (2002) Ghost3D: detecting body posture and parts using stereo. In: Motion and Video Computing, 2002. Proceedings. Workshop on, p 175180
[31] Hu M, Hu W, Tan T (2004) Tracking people through occlusions. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, pp 724– 727
[32] Hu W, Hu M, Zhou X, Tan T, Lou J, Maybank S (2006) Principal axis-based correspondence between multiple cameras for people tracking. Pattern Analysis and Machine Intelligence, IEEE Transactions on 28(4):663 –671, DOI 10.1109/TPAMI.2006.80
[33] Jain AK, Dass SC, Nandakumar K (2004) Soft biometric traits for personal recognition systems. In: International Conference on Biometric Authentication, Hong Kong, pp 731–738
[34] Jojic N, Frey B, Kannan A (2003) Epitomic analysis of appearance and shape. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, pp 34 –41 vol.1, DOI 10.1109/ICCV.2003.1238311
[35] Kim BS, Lee GG, Yoon JY, Kim JJ, Kim WY (2008) A method of counting pedestrians in crowded scenes. In: ICIC 08, Springer-Verlag, Berlin, Heidelberg, pp 1117–1126, DOI http://dx.doi.org/10.1007/978-3-540-85984-0_134
[36] Krahnstoever N, Rittscher J, Tu P, Chean K, Tomlinson T (2005) Activity recognition using visual tracking and RFID. In: Applications of Computer Vision and the IEEE Workshop on Motion and Video Computing, IEEE Workshop on, IEEE Computer Society, Los Alamitos, CA, USA, vol 1, pp 494–500, DOI http://doi.ieeecomputersociety.org/10.1109/ACVMOT.2005.17
[37] Lazebnik S, Schmid C, Ponce J (2005) A sparse texture representation using local affine regions. IEEE Transactions on Pattern Analysis and Machine Intelligence 27:1265–1278, DOI http://doi.ieeecomputersociety.org/10.1109/TPAMI.2005.151
[38] Lempitsky V, Zisserman A (2010) Learning to count objects in images. In: Advances in Neural Information Processing Systems
[39] Lin Z, Davis LS (2008) Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. ISVC pp 23–34
[40] Liu X, Krahnstoever N, Yu T, Tu P (2007) What are customers looking at?
[41] Lu J, Tan YP (2010) Gait-based human age estimation. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, pp 1718—1721
[42] Lu X, Jain AK (2004) Ethnicity identification from face images. In: Proc. SPIE, vol 5404, pp 114—123
[43] Macrae CN, Bodenhausen GV (2000) Social cognition: Thinking categorically about others. Annual Review of Psychology 51(1):93–120
[44] Marcialis G, Roli F, Muntoni D (2009) Group-specific face verification using soft biometrics. Journal of Visual Languages and Computing 20(2):101–109
[45] Nakajima C, Pontil M, Heisele B, Poggio T (2003) Full-body person recognition system. Pattern recognition 36(9):1997–2006
[46] Niinuma K, Unsang P, Jain AK (2010) Soft biometric traits for continuous user authentication. Information Forensics and Security, IEEE Transactions on 5(4):771–780
[47] Park U, Jain A (2010) Face matching and retrieval using soft biometrics. IEEE Transactions on Information Forensics and Security 5(3):406 –415, DOI 10.1109/TIFS.2010.2049842

[48] Park U, Jain A, Kitahara I, Kogure K, Hagita N (2006) Vise: Visual search engine using multiple networked cameras. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, vol 3, pp 1204 –1207, DOI 10.1109/ICPR.2006.1176

[49] Popa M, Rothkrantz L, Yang Z, Wiggers P, Braspenning R, Shan C (????) Analysis of shopping behavior based on surveillance system. In: Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on, IEEE, pp 2512–2519

[50] Popa M, Rothkrantz L, Wiggers P (2010) Products appreciation by facial expressions analysis. In: Proceedings of the 11th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing on International Conference on Computer Systems and Technologies, ACM, pp 293–298

[51] Prosser B, Zheng WS, Gong S, Xiang T (2010) Person re-identification by support vector ranking. In: Proceedings of the British Machine Vision Conference, BMVA Press, pp 21.1–21.11, DOI doi:10.5244/C.24.21

[52] Qi Y, Huang GC, Wang YH (2007) Carrying object detection and tracking based on body main axis. In: Wavelet Analysis and Pattern Recognition, 2007. ICWAPR '07. International Conference on, vol 3, pp 1237 –1240, DOI 10.1109/ICWAPR.2007.4421623

[53] Ran Y, Rosenbush G, Zheng Q (2008) Computational approaches for real-time extraction of soft biometrics. In: IEEE Int. Conf. On Pattern Recognition, pp 1–4

[54] Rodriguez M, Ali S, Kanade T (2009) Tracking in unstructured crowded scenes. In: Computer Vision, 2009 IEEE 12th International Conference on, pp 1389–1396

[55] Ryan D, Denman S, Fookes C, Sridharan S (2010) Crowd counting using group tracking and local features. In: 7th IEEE International Conference on Advanced Video and Signal-Based Surveillance, Boston, USA, pp 218–224

[56] Samangooei S, Guo B, Nixon M (2008) The use of semantic human description as a soft biometric. In: 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, 2008. BTAS 2008., pp 1 –7, DOI 10.1109/BTAS.2008.4699354

[57] Schwartz WR, Davis LS (2009) Learning discriminative appearance-based models using partial least squares. In: Computer Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on, pp 322–329

[58] Senior A, Brown L, Hampapur A, Shu C, Zhai Y, Feris R, Tian Y, Borger S, Carlson C (2007) Video analytics for retail. In: Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on, pp 423 –428, DOI 10.1109/AVSS.2007.4425348

[59] Shan C, Gong S, McOwan PW (2007) Learning gender from human gaits and faces. In: IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), 2007., pp 505–510, DOI 10.1109/AVSS.2007.4425362

[60] Stauffer C (2003) Estimating tracking sources and sinks. In: Event Mining Workshop, Madison, WI

[61] Tsai RY (1986) An efficient and accurate camera calibration technique for 3d machine vision. In: IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, pp 364–374

[62] Vaquero D, Feris R, Tran D, Brown L, Hampapur A, Turk M (2009) Attribute-based people search in surveillance environments. In: 2009 Workshop on Applications of Computer Vision (WACV), pp 1 –8, DOI 10.1109/WACV.2009.5403131

[63] Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: CVPR

[64] Wang L, Tan T, Hu W, Ning H (2003) Automatic gait recognition based on statistical shape analysis. IEEE transactions on image processing 12(9):1120–1131

[65] Wayman JL (1997) Large-scale civilian biometric Systems-Issues and feasibility. In: Proceedings of Card Tech/Secur Tech ID, vol 732

[66] Xiang-tao C, Zhi-hui F, Hui W, Zhe-qing L (2010) Automatic gait recognition using kernel principal component analysis. In: ICBECS, 2010 International Conference on, pp 1–4

[67] Xiaogang W, Doretto G, Sebastian T, Rittscher J, Tu P (2007) Shape and appearance context modeling. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pp 1–8

[68] Xu J, Denman S, Fookes C, Sridaharan S (2011) Activity modelling in crowded environments: A soft-decision approach. In: International Conference on Digital Image Computing: Techniques and Applications, Noosa Heads, Queensland, Australia
[69] Yu S, Tan T, Huang K, Jia K, Wu X (2009) A study on Gait-Based gender classification. IEEE Transactions on Image Processing 18(8):1905–1910, DOI 10.1109/TIP.2009.2020535
[70] Zhang E, Zhao Y, Xiong W (2010) Active energy image plus 2dlpp for gait recognition. Signal Processing 90(7):2295 – 2302
[71] Zheng WS, Gong S, Xiang T (2009) Associating groups of people. In: BMVC'09
[72] Zheng WS, Gong S, Xiang T (2011) Person re-identification by probabilistic relative distance comparison. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp 649–656