Person Reidentification Using Spatiotemporal Appearance

Niloofar Gheissari
National ICT
Canberra, Australia

Thomas B. Sebastian, Peter H. Tu, Jens Rittscher
GE Global Research
Niskayuna, NY, USA

Richard Hartley
Australian National University
Canberra, Australia

Abstract

In many surveillance applications it is desirable to determine if a given individual has been previously observed over a network of cameras. This is the person reidentification problem. This paper focuses on reidentification algorithms that use the overall appearance of an individual as opposed to passive biometrics such as face and gait. Person reidentification approaches have two aspects: (i) establish correspondence between parts, and (ii) generate signatures that are invariant to variations in illumination, pose, and the dynamic appearance of clothing. A novel spatiotemporal segmentation algorithm is employed to generate salient edgels that are robust to changes in appearance of clothing. The invariant signatures are generated by combining normalized color and salient edgel histograms. Two approaches are proposed to generate correspondences: (i) a model based approach that fits an articulated model to each individual to establish a correspondence map, and (ii) an interest point operator approach that nominates a large number of potential correspondences which are evaluated using a region growing scheme. Finally, the approaches are evaluated on a 44 person database across 3 disparate views.

1. Introduction

Many applications require the ability to reidentify an individual across multiple disjoint fields of view. In this paper, we consider reidentification algorithms that rely on the overall appearance of the individual as opposed to ones that use passive biometrics such as face [16] and gait [22]. An appearance-based algorithm must deal with several challenges such as: different camera angles and illumination conditions, variation in pose and the rapidly changing appearance of loose or wrinkled clothing. We assume, however, that individuals do not change their clothing between sightings. This is a reasonable assumption for many applications such as airport and subway surveillance. In many person reidentification applications, temporal reasoning and spatial layout of the different cameras can be used for pruning the set of candidate matches. To test the limits of our appearance-based reidentification algorithms we do not consider such information in this paper.

Several approaches have been proposed where invariant signatures based on the global appearance of an individual are compared. In [7] a color histogram of the region below the face (found by a face detector) serves as the signature for comparison. See [15] for a related approach using clothing color descriptors. Recently, the brightness transfer functions between different cameras have been used to track individuals over multiple non-overlapping cameras [14, 8]. It has been shown that the brightness transfer functions lie in a low-dimensional subspace, and can be learnt using a set of corresponding calibration objects [8]. Reidentification is then achieved by comparing the adjusted color histograms.

In contrast to the global appearance based methods previously discussed, recent advances in object recognition have demonstrated that comparing multiple local signatures can be effective in exploiting spatial relationships and achieving some robustness with respect to variations in appearance [2, 9]. The key to this methodology is the ability to establish correspondences between objects. Two approaches that are successful in this regard are interest point operators [9, 18] and model fitting [24].

There are two aspects to the person reidentification problem. First, we need to establish correspondences, i.e., determine which parts of one image should be compared to which parts in the second image. Second, we need to generate invariant signatures for comparing the corresponding parts. We hypothesize that the ability to compute the correspondences and generate invariant local signatures will also result in improved person reidentification. In addition, the performance will be enhanced if variation due to articulation can be directly addressed.

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

1.1. Overview of Approach

In this paper we develop two person reidentification approaches which use interest operators and model fitting, respectively, for establishing spatial correspondences between individuals. We also develop a novel spatiotemporal segmentation that utilizes spatial and temporal cues to generate invariant signatures for clothing.

The responses of many interest point operators will not persist over extended periods of time due to the dynamic nature of the appearance of a person [11]. To address this issue, our strategy is to choose an operator that generates a large number of responses in regions with high information content, thus increasing the probability of establishing true correspondences between images of the same individual. In this paper, the Hessian affine invariant operator [11] is used for this purpose. Signature matching is used to establish correspondences between two sets of interest points. A match score is computed based on the cardinality of the final set of correspondences.

In contrast to the interest point operator approach which generates a large number of potential correspondences, model-based algorithms establish a mapping from one individual to another. In this paper we use a decomposable triangulated graph [1, 5] to model the articulated shape of a person. A dynamic-programming algorithm is used to fit the model to the image of the person [1, 5]. Model fitting localizes different body parts such as arms, torso, legs and head, thus facilitating the comparison of appearance and structure between corresponding body parts.

We now describe how the invariant signatures for comparing different regions are generated by combining color and structural information. The color information is captured by histograms based on hue and saturation. Some invariance to differences in ambient illumination is achieved via normalization. Unlike most rigid objects, the structural appearance of loose fitting or wrinkled clothing on perambulating individuals is highly dynamic. Hence, the application of a traditional edge operator [4] will produce many spurious edges corresponding to wrinkles and folds in clothing. To address this issue, a novel spatiotemporal segmentation algorithm that generates salient edgel information is applied to the imagery. The watershed algorithm is used to generate an over-segmentation of each frame. A spatiotemporal graph is then generated by treating each region as a node and placing edges between spatially and temporally adjacent regions. A graph partitioning algorithm that models each cluster as a minimum spanning tree is then used to generate salient edgels corresponding to the boundaries of each type of clothing. Finally, the region signatures are then augmented with local histograms of these salient edgels.

The paper is organized as follows. In Section 2 the local signature generation is described. In Section 3 the spatiotemporal segmentation algorithm is presented. Sections 4 and 5 describe the interest operator and model fitting approaches to correspondence generation. In Section 6 the algorithms are evaluated against a database of 44 individuals observed over 3 disparate views. Section 7 discusses the two approaches and concludes the paper.

2. Signature Generation

This section describes how the signature or feature vector h_i is generated for given a local support region i. The first component of this feature vector is a histogram of the hue and saturation of the region. To overcome the sensitivity of HSV histograms to changes in illumination and shadows in outdoor scenes, we use the definition of hue which is invariant to brightness and Gamma [20]. This definition of hue for a given RGB color space is as follows:

$$H = \arccos \frac{\log(R) - log(G)}{\log(R) + \log(G) - 2\log(B)}, \qquad (1)$$

as compared to the traditional definition which is

$$H = \arctan \frac{0.5[(R-G) + (R-B)]}{\sqrt{(R-G)(R-G) + (R-B)(G-B)}}. \qquad (2)$$

The second part of h_i represents the structural qualities of the region. As in [9, 10] a histogram of edgels that are contained in the region i are used for this purpose. The direct application of traditional edge detection algorithms such as Canny [4] to images of clothing produces many spurious responses. Hence, salient edgels generated by a novel spatiotemporal segmentation algorithm is used in this application (see Section 3). The spatiotemporal segmentation rejects edge information that are temporally unstable . Only edgels that are interior to the foreground are considered. Each edgel encodes the dominant local boundary orientation (vertical or horizontal) as well as the ratios between the RGB color components of the two regions on either side of the edgel. The ratios of the RGB color components are each quantized to 4 possible values. This results in a 7 bit edgel representation.

In the matching process the distance between two signatures h_i and h_j (of n bins) is defined using the intersection histogram [20]

$$D(h_i, h_j) = 1 - 2\frac{\sum_{k=1}^{n} \min(h_i(k), h_j(k))}{\sum_{k=1}^{n} (h_i(k) + h_j(k))} \qquad (3)$$

3. Spatiotemporal Segmentation

In this section we propose an algorithm that produces stable structural information in the form of edgels which are used for defining region signatures. Even though many articles of clothing are derived from materials with uniform reflectance properties, a given type of material may appear

quite different across an image and over time. This is because the surface normals of loosely-fitting clothing under articulated motion are highly dynamic. The proposed spatiotemporal segmentation method groups pixels that belong to the same type of fabric. Salient edgels are those pixels that are on the boundaries between two such groupings.

Observe that intra-fabric boundaries are not stable over time due to folds and wrinkles. We exploit this idea in the spatiotemporal segmentation. For a given time window an over-segmentation is performed on each image. This results in a set of contiguous regions $\mathbf{R} = \{r_i^t\}$, where r_i^t is the i^{th} region of image t. A graph $\mathbf{G} = \{\mathbf{V}, \mathbf{E}\}$ is defined for a set of vertices $\mathbf{V} = \{v_i^t\}$ and edges $\mathbf{E} = \{e_{i,i'}^{t,t'}\}$ where v_i^t corresponds to region r_i^t and $e_{i,i'}^{t,t'}$ is an edge connecting vertices v_i^t and $v_{i'}^{t'}$. Region grouping is performed by partitioning $\mathbf{G}$ into a set of clusters. A number of authors [12, 13] have used region grouping over time for the purpose of foreground - background separation. However, our objective differs from these applications in that the goal is to achieve a stable segmentation of the foreground. In [25] segmentation that is consistent across neighboring frames is developed and is used for video editing.

The over-segmentation used to define $\mathbf{V}$ is performed in two stages. First, a Sobel operator is applied to the foreground of each grey level image and this is followed by Gaussian filtering. Second, a watershed segmentation algorithm [21] is applied. This results in regions of uniform intensity value for the foreground of the image. This method is appropriate for articles of clothing that are not overly textured.

Given $\mathbf{V}$, the edge structure can then be defined. Two types of edges are constructed: spatial and temporal. If two regions r_i^t and $r_{i'}^t$ share a common boundary, then a spatial edge $e_{i,i'}^{t,t}$ is formed. For each region r_i^t, the region $r_{\hat{i}}^{t+1}$ is determined such that $r_{\hat{i}}^{t+1}$ has the highest likelihood of corresponding to the same material as r_i^t. This establishes the temporal edge $e_{i,\hat{i}}^{t,t+1}$. The selection of $r_{\hat{i}}^{t+1}$, is determined based on estimates of the motion field. A frequency image $F_{N,t}(x,y)$ is defined as

$$F_{N,t}(x,y) = \sum_{k=0}^{N} H(I_t(x,y) - I_{t+k}(x,y)) \tag{4}$$

where $I_t(x,y)$ is the intensity value of pixel (x,y) at time t and

$$\begin{array}{ll} H(z) = 1 & if|z| < \delta \\ H(z) = 0 & otherwise \end{array} \tag{5}$$

for a threshold δ. For a given region with uniform intensity and uniform motion, the values of $F_{N,t}(i,j)$ will be higher on the side of the region that corresponds to the direction of forward motion. For each overlapping region $r_{i'}^{t+1}$, we compute the integral of $F_{N,t}(i,j)$ over the intersection of r_i^t and $r_{i'}^{t+1}$. The overlapping region with the highest frequency integral is selected for a temporal edge. See Figure 1 for an example of the frequency image.

If two adjacent regions correspond to the same piece of fabric they will periodically have a similar appearance. This is because intra-fabric boundaries are inherently unstable. Based on this, the edge weight $w_{i,i'}^{t,t'}$ is now defined as the cost of grouping two regions together:

$$w_{i,i'}^{t,t} = |M(i,t) - M(i',t)| \tag{6}$$

$$w_{i,i'}^{t,t+1} = \frac{1}{3}|M(i,t) - M(i',t+1)| \tag{7}$$

where $M(i,t)$ is the median intensity value for region r_i^t. Note that a temporal edge allows for greater variation in appearance. We argue that two regions should be grouped if there is a low cost path connecting them through space, time or a combination of both. Thus our graph partitioning algorithm is based on a search for clusters that have low-cost spatiotemporal minimal spanning trees.

3.1. Graph Partitioning

Once the spatiotemporal graph $\mathbf{G}$ has been generated for ten consecutive frames, the graph partitioning algorithm proposed in [6] is used for grouping spatiotemporally similar regions. The basic idea of this partitioning algorithm is to merge connected clusters whenever the distance between them is less than the internal variation of each of the individual clusters. To efficiently implement this approach each cluster $\mathbf{C}$ is represented by the minimum spanning tree $\mathbf{E^C}$ passing through all its vertices $\mathbf{V^C}$. The maximum edge weight of the minimum spanning tree is used to define the internal variation of the cluster $\mathbf{C}$, i.e.,

$$I(\mathbf{C}) = \max(w_{i,i'}^{t,t'}) s.t.\ e_{i,i'}^{t,t'} \in \mathbf{E^C} \tag{8}$$

Given two clusters $\mathbf{C_m}$ and $\mathbf{C_n}$, the inter-cluster distance $D(\mathbf{C_m}, \mathbf{C_n})$ is defined as the lowest edge weight between them, i.e.,

$$\begin{array}{l} D(\mathbf{C_m}, \mathbf{C_n}) = \min(w_{i,i'}^{t,t'}) \\ s.t.\ v_i^t \in \mathbf{V^{C_m}}, v_{i'}^{t'} \in \mathbf{V^{C_n}}, e_{i,i'}^{t,t'} \in \mathbf{E}. \end{array} \tag{9}$$

Two clusters are merged if the inter-cluster distance is small when compared to the internal variation of the individual clusters. Specifically, clusters $\mathbf{C_m}$ and $\mathbf{C_n}$ are merged if

$$D(\mathbf{C_m}, \mathbf{C_n}) \leq MI(\mathbf{C_m}, \mathbf{C_n}) \tag{10}$$

where

$$MI(\mathbf{C_m}, \mathbf{C_n}) = \min(I(\mathbf{C_m}) + \kappa/|\mathbf{C_m}|, I(\mathbf{C_n}) + \kappa/|\mathbf{C_n}|). \tag{11}$$

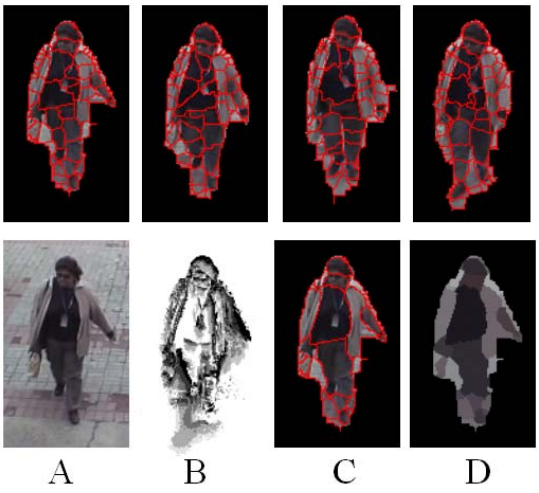

Figure 1. Upper row: over segmentation for frames 0,3,6,9. Lower row: A) original image, B) Frequency image for N=10, C) Final segmentation after graph partitioning, D) median image for final segmentation

The factor $\kappa/|\mathbf{C}|$ is based on the size of the cluster and encourages the formation of larger clusters. Otherwise, in the beginning of the merging process, the internal variation of very small clusters tends to be too small, causing the merging process to stop prematurely.

A greedy algorithm is proposed in [6] to obtain the graph segmentation that satisfies the above conditions. All edges in the spatiotemporal graph are sorted according to non-decreasing edge weights, and are then processed in that order. Let an edge $e_{i,i'}^{t,t'}$ between two separate clusters $\mathbf{C_m}$ and $\mathbf{C_n}$ be the one under consideration. If $w_{i,i'}^{t,t'} \leq MI(\mathbf{C_m}, \mathbf{C_n})$, then $\mathbf{C_m}$ and $\mathbf{C_n}$ are merged and the edge $e_{i,i'}^{t,t'}$ added to the minimum spanning tree of the combined cluster. This step is repeated until all edges have been processed. It has been shown in [6] that the segmentation produced by the above algorithm is optimal in that the maximum edge weight for the minimum spanning tree for each cluster is smaller than the weights of all edges to each of their neighboring clusters. See Figure 1 for an example of the application of this algorithm.

3.2. Foreground-Background Separation

While foreground background segmentation is not the focus of this paper, estimates of the foreground patches are required. For this application we have found that a reasonable approach is to first compute the maximum frequency image which is defined as :

$$MF_{N,t}(i,j) = max(F_{N,t}(i,j), F_{-N,t}(i,j)) \qquad (12)$$

This results is low values near motion boundaries, after thresholding and morphological filtering (to fill in the centers), reasonable foreground/background segmentations are achieved.

4. Interest-Point Matching

This section describes how interest operators are used to establish correspondences between individuals. Given an image of a person, the Hessian Affine invariant interest operator [11] is used to nominate points of interest. The operator is limited to foreground patches extracted using mechanisms described in section 3.2. The Hessian operator is not stable over time. However, when compared to other methods [11], it provides a large number of interest points and it is more informative with respect to color variation. This increases the probability of generating true correspondences between images of the same individual. For each interest point i, a feature vector h_i is generated (see section 2) based on a circular support region *C(i,r)* of fixed radius r centered at position i. In order to limit the influence of foreground segmentation errors, interest points that contain large amounts of background are not considered.

When two images I and J are compared, an initial set of correspondences are nominated. The merit of a potential match $(i \rightarrow j)$ is evaluated using equation 3. Inverse matching is used to ensure consistency of the correspondences. For each interest point i in image I, the most likely correspondence i' in image J is determined. If the distance between the signatures of i and i' is below a threshold, then the most likely interest point i'' in image I corresponding to point i' is determined. If the Euclidean distance between i and i'' is smaller than a threshold, then the correspondence $(i \rightarrow i')$ is accepted.

A final validation stage is used to prune the initial correspondences. The support regions for corresponding interest points are expanded iteratively in the vertical direction. We again construct the region signatures for each expanded region and compare them using equation 3. The process continues for a fixed number of iterations or there is too much overlap with the background. If the distance between the two signatures remains below a fixed threshold, the correspondence is accepted into the final set of correspondences.

The score given to the match between images I and J is based on the cardinality of the final set of correspondences. Two examples of this matching process are shown in Figure 2. The efficacy of this matching algorithm is evaluated in section 6.

5. Dynamic-Programming Model Fitting

In contrast to the interest operator algorithm, we now consider a model-based approach that generates a correspondence between different body parts such as the head, arms, legs and torso. In other words, we we need to match the torso of an individual in one scene with the torso in the

Figure 2. Two examples of the interest operator matching algorithm. On the left are two images that are to be compared. On the right are the identified correspondences and associated signature regions.

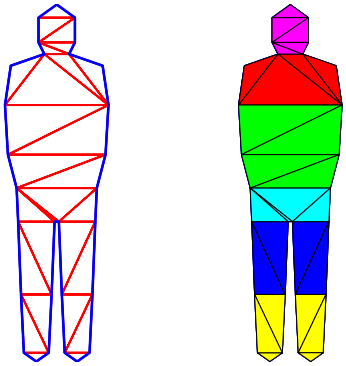

Figure 3. Left: An example of a decomposable triangulated graph used as a person model. The solid (blue) edges correspond to the boundary of the person while the light (red) edges are interior edges. Note that in this work we consider models without arms, mostly because most individuals in the database, have their arms next to their torso. Right: Partitioning of the person used for generating signatures for comparison.

second scene and so on. This presents a significant challenge as the relative location of arms, legs and torso of an individual varies from one scene to another. To address this problem we propose to use a model-based top-down segmentation of an individual in a scene where the different parts are accurately localized. This segmentation is used to establish the correspondence between the different body parts of an individual in two scenes, which facilitates a comparison of the appearance of these body parts.

We use a decomposable triangulated graph as a novel method for model fitting to people. See Figure 3 for an example. Several researchers have used decomposable triangulated graphs to represent deformable shapes [1, 19, 5, 23]. These graphs are a collection of cliques of size three and have a perfect elimination order for their vertices, i.e., there exists an elimination order for all vertices such that (i) each eliminated vertex belongs only to one triangle, and (ii) a decomposable triangulated graph results from eliminating the vertex. As these graphs support a perfect elimination order, the model optimization can be efficiently done using a dynamic programming algorithm [1].

We now describe how the decomposable triangulated graph is used for modeling and segmenting people in a scene using an energy minimization approach. The starting point for the model-fitting algorithm is the bounding box of the person of interest. Let the model be a decomposable triangulated graph T with n triangles, $T_i, i = 1, \ldots, n$. We are seeking a function g that maps the model to the image domain such that the consistency of the model with salient image features is maximized, and deformations of the underlying model is minimized. This function g is restricted to being a piecewise affine map [5], where the deformation of each triangle $g_i(T_i)$ in the model is an affine transformation. The energy functional $E(g, I)$ that we are trying to minimize can then be written as a sum of costs with one set of costs for each triangle in the model. Specifically,

$$\begin{aligned} E(g,I) &= \textstyle\sum_i E_i(g_i, I) \\ &= \textstyle\sum_i E_i^{data}(g_i, I) + E_i^{shape}(g_i) \end{aligned} \tag{13}$$

where I denotes the underlying image features.

We now discuss how the shape and data costs for each triangle in the model are formulated. The shape cost for each triangle is defined in terms of the polar decomposition of its affine transformation [17]. For an affine matrix A the polar decomposition is defined as

$$\begin{aligned} A &= \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} s_x & s_h \\ s_h & s_y \end{bmatrix} \\ &= R(\psi)S. \end{aligned}$$

S is the scale-shear matrix and $R(\psi)$ is the closest possible rotation matrix to A[1]. The polar decomposition can be derived from the commonly used singular value decomposition of the affine transformation. The shape term is then defined as

$$E^{shape} = \log(\frac{\lambda_1}{\lambda_2})^2 + \log(1 + s_h)^2 \tag{14}$$

where λ_1 and λ_2 are eigenvalues of the scale-shear matrix. The first term is the log-anisotropy term and penalizes changes in height-width ratio [3, 5], while the second term penalizes shear.

[1]The closeness is measured using Frobenuis matrix norm, i.e., $R(\psi) = argmin_Q \parallel Q - A \parallel_F$, where$Q^TQ = Iand \parallel Q - A \parallel_F = \sum_{i,j}(q_{ij} - m_{ij})^2$.

Figure 4. This figure illustrates two examples of fitting the decomposable triangulated model to individuals. The cropped image image, edge feature image, foreground mask, and fitting results are shown from left to right. The green dots show the candidate locations for the model points. Observe that the model fits well to the individuals despite the presence of bags, shadows, additional interior edges due to different clothing etc.

The data cost in the energy functional attracts the model to salient image features. Note that the decomposable triangulated graph has both boundary edges and interior edges, and that the data costs are defined only for boundary edges. The data cost for all interior edges is zero. Two complementary sets of image features are used to define the data cost: salient edges in the image which are detected using Canny's algorithm [4] and a foreground mask that is obtained using the spatiotemporal segmentation discussed in Section 3. Note that the model fitting approach is less sensitive to the presence of spurious edges than to missing ones, hence we use a combination of Canny edges and spatiotemporal foreground mask to compute the data cost.

After the edges are extracted using Canny's algorithm, a Euclidean distance transform is applied to get the edge feature image D. The edge cost measures how far a boundary triangle edge is from Canny edges. The average edge feature value along the sampled triangle edge $L = (x_i, y_i), i = 1, \ldots, n$ is used as the edge cost, i.e.,

$$E^{edge} = \frac{1}{n} \sum_{i=1}^{n} D(x_i, y_i). \tag{15}$$

The foreground cost measures consistency of the model with the foreground mask and is defined by the relative number of foreground pixels in a window on either side of the boundary triangle edge.

$$E^{fg} = 1 - \left| \frac{N_1^{fg}}{N_1} - \frac{N_2^{fg}}{N_2} \right| \tag{16}$$

where N_1^{fg} and N_1 are the number of foreground pixels and total number of pixels on one side of the window. N_2^{fg} and N_2 are similarly defined for the other side. Note that this term is small when the boundary edge is along the foreground mask.

The dynamic-programming algorithm for computing the optimal deformation of the model is now described. We want to find g that maps the vertices of the model to image locations such that the energy functional in Equation 13 is minimized. The dynamic programming does an exhaustive search of the candidate locations to find the global optimum. We restrict the candidate locations for the vertices of the model to be the boundary of the foreground mask and Canny edges. Since the triangulated model used here has a perfect elimination order and the cost defined in Equation 13 is extensible, a serial dynamic-programming algorithm [1, 5] can be used for optimization. At each iteration of the algorithm, the perfect elimination order is used to eliminate one vertex from the model, and its optimal location is encoded in terms of its two adjacent vertices. This process is repeated until all vertices are eliminated. The final location of all vertices in the model is computed by standard backtracking. Figure 4 shows the results for two representative cases.

Once the model fitting is done, the appearance and shape signature for an individual is computed as follows. The individual is partitioned into salient body parts using the fitted model as is illustrated in Figure 3. As an example, consider how the signature is generated for the upper torso. Using all the triangles that correspond to the upper torso (colored red in Figure 3) the appearance and shape signature is extracted. Similarly, the signatures for all salient body parts are computed and is compared using the Equation 3.

6. Experimental Results

For purposes of evaluation, 44 different individuals were recorded using three different cameras with disjoint views. See Figure 6. The subjects were recorded entering a corporate campus and were in no way coached or rehearsed. Two to four key frames were selected for each person from each camera. Each image in the database is indexed as I_{pfc} where p encodes the person id, f encodes the key frame number and c encodes the camera view. All the reidentification algorithms are evaluated in the following manner:

- Each image I_{pfc} is compared against the set of images $(I_{pfc'})$ such that $c \neq c'$.

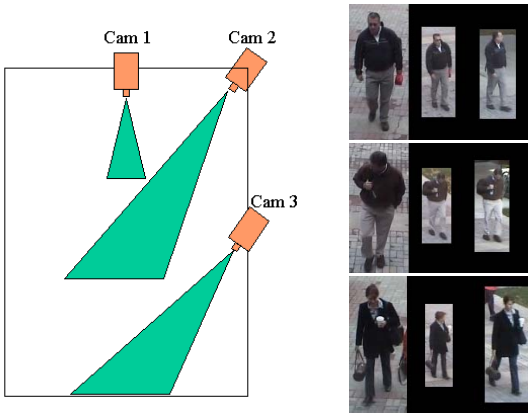

Figure 5. Left: The layout of cameras used for collecting the data for the experiments. Camera 2 is placed roughly at twice the height of Cameras 1 and 3. Right: The representative samples of key frames for three individuals from all three camera views. The person bounding boxes are placed on a black background to illustrate the differences in resolution in the different views.

- For each person/camera combination, the maximum ranking true match for all it's key frames is determined.
- The number of times that a maximum ranking true match is higher than a given value is then tabulated.

This evaluation scheme is analogous to a standard surveillance scenario where an operator would query a person reidentification system with multiple images of the same individual captured over a short period of time from a particular camera. Any hits from these queries would result in a success.

Three reidentification algorithms are evaluated in this manner. The first two algorithms use the interest operator (see Section 4) and model fitting (see Section 5) approaches for generating correspondences. A third algorithm, referred to as the bounding box method, computes a single signature using the foreground pixels in the bounding box of each individual and calculates the distance between the resulting monolithic feature vectors to perform the matching. This serves as a baseline implementation for comparison.

Figure 6 reports the performance of the three algorithms. The model fitting approach results in the best performance with approximately 60 percent of the queries achieving a top ranking true match and over 90 percent of the queries generating a true match in the top ten. The interest operator method achieves a top ranking true match 25 percent of the time and a true match is found in the top ten 65 percent of the time. It should be noted that the performance of this approach may improve with higher image resolotution. The performance of the bounding box approach is comparable to that of the interest point approach. Figure 6 shows the top ranking images for a number of queries using the model fitting algorithm. See Section 7 for further discussion of these different approaches.

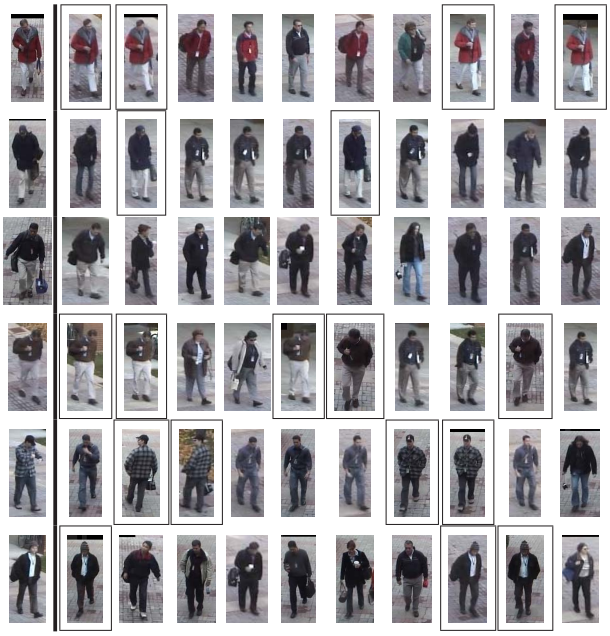

Table 1. Top ten matches using the model-based algorithm. The query image is shown in the left column, and the remaining columns are the top matches ordered from left to right. A box is used to highlight when a match corresponds to query. Third row shows an example where the correct match is not present in the top ten matches.

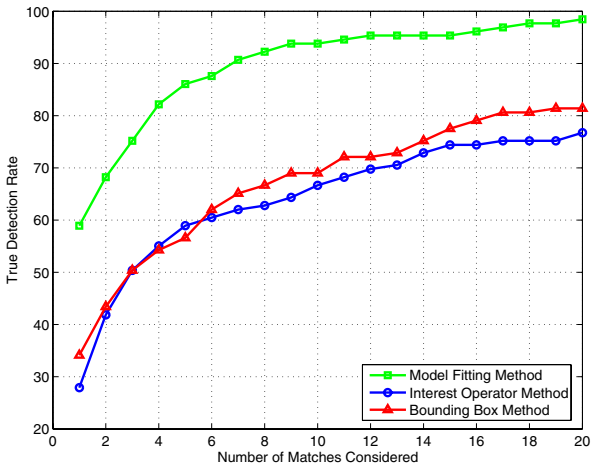

Figure 6. This figure compares the overall performance of the interest operator, model fitting, and bounding box approaches. The percent correct detection rate is plotted vs. the number of matches considered. Note that the model fitting approach is the best performer, and the performance of bounding box approach and interest point approach is comparable.

7. Conclusion

This paper has demonstrated that by both establishing accurate correspondences and generating invariant signatures, greatly improved person reidentification can be achieved. This confirms the hypothesis put forth in the introduction. This paper has presented a novel application of triangulated model fitting to people that directly addresses issues associated with articulation. Even though human subjects are highly deformable, an interest operator approach to correspondence generation was demonstrated. In addition a new spatiotemporal segmentation algorithm has been developed that provides structural information that is invariant to the dynamic appearance of clothing.

This work can be extended in the following directions. First, observe that the model fitting approach generates an ordered set of signatures, enabling efficient indexing schemes for database queries. This would be valuable for forensic applications involving large numbers of cameras capturing imagery over extended periods of time. Finally, our experiments indicate that the interest operator and model fitting approaches can be combined to form a hierarchical person reidentification approach. In the first stage, an interest point operator algorithm is used as a fast way for reducing the number of candidate matches. In the second stage, a model-fitting approach will be applied for cases that confound the interest operator method. We plan to evaluate this hierarchical approach using a much larger database.

References

[1] Y. Amit and A. Kong. Graphical templates for model registration. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(3):225–236, 1996. 2, 5, 6

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(4):509 – 522, 2002. 1

[3] F. L. Bookstein. Size and shape spaces for landmark data in two dimensions. *Statistical Science*, 1(2):181–242, 1986. 5

[4] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Analysis Machine Intelligence*, 8(6):679–698, 1986. 2, 6

[5] P. F. Felzenszwalb. Representation and detection of deformable shapes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(2):208–220, 2005. 2, 5, 6

[6] P. F. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004. 3, 4

[7] G. Jaffre and P. Joly. Costume: A new feature for automatic video content indexing. In *Proceedings of RIAO*, pages 314–325, 2004. 1

[8] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages II:26–33, 2005. 1

[9] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, page 1150, 1999. 1, 2

[10] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision,*, pages I:69–82. Springer, 2004. 2

[11] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, 2005. 2, 4

[12] F. Moscheni, S. Bhattacharjee, and M. Kunt. Spatiotemporal segmentation based on region merging. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(9):897–915, September 1998. 3

[13] L. Patras, E. Hendriks, and R. Lagendijk. Video segmentation by MAP labeling of watershed segments. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(3):326–332, 2001. 3

[14] F. Porikli. Inter-camera colour calibration using cross-correlation model function. In *International Conference on Image Processing*, pages II:133–136, 2003. 1

[15] J. Seigneur, D. Solis, and F. Shevlin. Ambient intelligence through image retrieval. In *Conference on Image and Video Retrieval*, pages 526–534. Springer, 2004. 1

[16] A. Senior, R.-L. Hsu, M. A. Mottaleb, and A. K. Jain. Face detection in color images. *IEEE Trans. Pattern Analysis Machine Intelligence*, 24(5):696–706, 2002. 1

[17] K. Shoemake and T. Duff. Matrix animation and polar decomposition. In *Proceeding of Graphics Interface*, pages 258–264, 1992. 5

[18] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, 2003. 1

[19] Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(7):814–827, 2003. 5

[20] M. J. Swain and D. H. Ballard. Indexing via color histograms. In *DARPA Image Understanding Workshop*, 1990. 2

[21] L. Vincent and P. Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 13(6):583–598, 1991. 3

[22] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(12):1505–1518, 2003. 1

[23] J. Zhang, R. Collins, and Y. Liu. Representation and matching of articulated shapes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages II:342–349, 2003. 5

[24] T. Zhao and R. Nevatia. Car detection in low resolution aerial images. *Image and Vision Computing*, 21(8):693–703, 2003. 1

[25] C. L. Zitnick, N. Jojic, and S. B. Kang. Consistent segmentation for optical flow estimation. In *International Conference on Computer Vision*, pages 1308–1315, 2005. 3