Gait Recognition using Image Self-Similarity

Chiraz BenAbdelkader, Ross G. Cutler, and Larry S. Davis

Abstract

Gait is one of the few biometrics that can be measured at a distance which makes it useful for passive surveillance as well as biometric applications. Gait recognition research is still at its infancy, however, and we have yet to solve the fundamental issue of finding gait features which at once have sufficient discrimination power and can be extracted robustly and accurately from low-resolution video. This paper describes a novel gait recognition technique that uses the image self-similarity of a walking person. We contend that the similarity plot encodes a projection of gait dynamics. It is also correspondence-free, robust to segmentation noise, and works well with low-resolution video. We test the method on multiple data sets of varying sizes and degrees of difficulty. Performance is best for fronto-parallel viewpoints. A recognition rate of 98% is achieved for a fronto-parallel data set of 6 people, and 70% for a fronto-parallel data set of 54 people.

Keywords

Gait recognition, Human identification at a distance, Human movement analysis, Behavioral biometrics, Pattern recognition.

I. Introduction

A. Motivation

Gait is a relatively new and emergent behavioral biometric [1], [2] that pertains to the use of an individual's walking style (or 'the way they walk') to determine identity. *Gait recognition* is the term typically used in the computer vision community to refer to the automatic extraction of visual cues that characterize the motion of a walking person in video and using for identification purposes. Gait is particularly an attractive modality for passive surveillance since, unlike most biometrics, it can be measured at a distance, hence not requiring interaction with or cooperation of the subject. However, gait features exhibit a high degree of intra-person variability, being dependent on various physiological, psychological and external factors such as footwear, clothing, surface of walking, mood, illness, fatigue, etc. The question then arises as to whether there is sufficient gait variability *between* people that can discriminate them even in the presence of large variation *within* each individual.

There is indeed strong evidence originating from psychophysical experiments [3], [4], [5] and gait analysis research (a well-advanced multi-disciplinary field that spans kinesiology, physiotherapy, orthopedic surgery, ergonomics, etc.) [6], [7], [8], [9], [10] that gait dynamics contain a signature that is characteristic of, and possibly unique to, each individual.

From a biomechanics standpoint, human gait consists of synchronized, integrated movements of hundreds of muscles and joints of the body. These movements follow the same basic bipedal pattern for all humans, and yet vary from one individual to another in certain details (such as their relative timing and magnitudes) as a function of their entire musculo-skeletal structure (i.e. body mass, limb lengths, bone structure, etc.). Because this structure is difficult to replicate, gait is believed to be unique to each individual and can be completely characterized by a few hundred kinematic parameters, namely the angular velocities and accelerations at certain joints

Chiraz BenAbdelkader: Identix Corporation, One Exchange Place, Jersey City NJ 07302 USA. Email: chiraz@cs.umd.edu. Ross Cutler: Microsoft Research, One Microsoft Way, Redmond, WA 98052-6399, USA. Email: rcutler@microsoft.com. Larry Davis: Department of Computer Science at the University of Maryland, College Park, MD 20742 USA. Email: lsd@umiacs.umd.edu.

and body landmarks [6], [7]. Achieving such a complete characterization *automatically* from low-resolution video remains an open research problem in computer vision. The difficulty lies in that feature detection and tracking is error prone due to self-occlusions, insufficient texture, etc. This is why computer-aided motion analysis systems still rely on special wearable instruments (such as LED markers) and walking surfaces [9].

Luckily, we may not need to recover 3D kinematics for gait recognition after all. In Johansson's early psychophysical experiments [3], human subjects were able to recognize the type of movement solely by observing light bulbs attached to a few joints of the moving person. The experiments were filmed in total darkness so that only the bulbs, a.k.a. Moving Light Displays (MLD's), are visible. Similar experiments later suggested that the identity of a familiar person ('a friend') [5], as well as the gender of the person [4], may be recognizable from their MLDs. While it is widely agreed that these experiments provide evidence about motion perception in humans, there is no consensus on how the human visual system actually interprets this MLD-type stimuli. Two main theories exist: The first maintains that people recover the 3D structure of the moving object (person) and subsequently use it for recognition. The second theory states that motion information is directly used for recognition, without structure recovery in the interim [11]. This seems to suggest that the raw spatiotemporal (XYT) patterns generated by the person's motion in an MLD video encode information that is sufficient to recognize their movement.

In this paper, we describe a novel gait recognition technique that derives classification features directly from these spatiotemporal patterns. Specifically, it computes the image self-similarity plot (SSP), defined as the correlation of all pairs of images in the sequence. Normalized feature vectors are extracted from the SSP and used for recognition. Related work has demonstrated the effective use of SSP's in recognizing different types of biological periodic motions, such as those of humans and dogs, and applied the technique for human detection in video [12]. We use them here to classify the movement patterns of different people. We contend that the SSP encodes a projection of planar gait dynamics and hence a 2D signature of gait. Whether it contains sufficient discriminant power for accurate recognition is what we set to determine.

It is our belief that a successful gait recognition method should at the very least be able to work well both with low-resolution (surveillance) video and under normal levels of individual variability. Gait recognition is still at its infancy. While several methods have been suggested in the recent past, they are either not robust enough or have not been tested on realistic data that is sufficiently representative of gait variation. The method of this paper is a first step to meeting these minimum requirements. The computation of the SSP is correspondence-free, is robust to segmentation noise and can be done with fairly low-resolution images. The method is view-dependent (since it is inherently appearance-based), however this is circumvented via view-based recognition. We assess the performance of the method on several data sets of varying degrees of difficulty, including a large (surveillance-quality) outdoor data set of 54 people, and a multi-view data set of 12 people taken from 8 viewpoints.

B. Assumptions

The method makes the following assumptions:

- People walk with constant velocity for about 3-4 seconds.
- People are located sufficiently far from the camera.
- The frame rate is greater than twice the frequency of the walking.
- The camera is stationary.

C. Organization of the Paper

The rest of the paper is organized as follows. Section II discusses related work in the computer vision literature and Section III describes the method in detail. We assess the performance of the method on a number of different data sets in Section IV, and finally conclude in Section V.

II. Related Work

The interest in gait recognition is best evidenced by the near-exponential growth of the size of related literature over the past few years [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [19], [25], [26], [27], [28], [29], [30], [31], [32]. Gait recognition is generally related to human movement analysis methods that automatically detect and/or track human motion in video for a variety of applications- surveillance, video-conferencing, man-machine interfaces, smart rooms, etc. For good surveys on this topic, see [11], [33], [34]. It is perhaps most closely associated with the subset of methods that analyze whole-body movement, such as for human detection [35], [12], [36] and activity recognition [37], [38], [39], [40].

A basic characteristic in all of these methods is that they exploit motion as the primary cue for detection/recognition. They typically consist of two stages:

- A motion extraction stage, which derives motion information from the image sequence and organizes it into some compact form (or representation). This can be viewed as a data reduction step that keeps only information relevant to the recognition task at hand. The effectiveness or goodness of motion patterns is essentially defined by (1) how discriminative they are, i.e. whether we can separate different individuals based on these patterns alone, particularly in the presence of intra-person variability, and (2) how robust they are to measurement noise (due to segmentation, shadows, etc.).
- A recognition stage, which applies some standard pattern classification technique on the obtained motion patterns. As in any pattern classifier, this may involve a dimensionality reduction step (such as principal components analysis) prior to classification.

Most existing methods use a simple, K-nearest neighbor-like approach for the recognition stage, but differ widely in the type of motion pattern they use. In fact gait recognition research is still mostly focused on the first stage, and rightly so, since this is of a more fundamental importance in classifier design (than the second stage). There are two main approaches for extracting relevant motion information from images for gait recognition: holistic [14], [15], [16], [17], [18], [19], [23], [24], [25], [28], [29], [30], [31], [32] and feature-based [41], [20], [42], [21], [22], [27], [43], [44], [26]. The holistic vs. feature-based dichotomy can also be regarded as global vs. local, non-parametric vs. parametric, and pixel-based vs. geometric. This dichotomy is certainly recurrent in visual pattern recognition problems such as face recognition [45], [46]. In the sequel, we describe and critique examples from both approaches, and relate them to the two gait recognition techniques of this paper.

A. Holistic Approach

The holistic approach characterizes body movement by the statistics of the spatio-temporal patterns (XYT) generated in the image sequence by the walking person. Although typically these patterns have no direct physical meaning, intuitively they capture both the static and dynamic properties of body shape. There are many ways of extracting XYT patterns from the image sequence of a walking person. However, in a nutshell, they all either extract raw XYT data (namely the temporal sequence of binary/color silhouettes or optical flow images), or a mapping of this data to a more terse 1D or 2D signal.

Perhaps the simplest approach is to use the sequence of binary silhouettes spanning one gait cycle, scaled to a certain uniform size [15], [32]. A slight variation of this uses silhouettes that correspond to certain gait poses only, particularly those that can be easily detected, such as at the double-support and mid-stance poses [30]. Classification is achieved either by directly comparing (correlating) these silhouette sequences [32], [30], or by first mapping them to a reduced feature space, obtained via some dimensionality reduction technique (namely PCA), then comparing them in this space [15]. Decent classification rates (above 90%) are reported on many data sets. However, these methods seem to be the most sensitive (among holistic methods) to appearance changes particularly due to clothing style and camera viewpoint. Their ability to perform as well on noisy low resolution video seems unlikely. Nonetheless, these methods provide good baseline performance against which to evaluate other more contrived gait recognition methods.

Instead of using the entire silhouette, other methods use a signature of the silhouette, such as binary shape moments, and vertical (row) or horizontal (column) projection histograms [14], [18], [28], [30], [31]. They essentially collapse the XYT data into a more terse 1D or 2D signal(s). Niyogi and Adelson [14] use snakes to track and extract four (two-dimensional) XT sheets that encode the person's inner and outer bounding contours. Recognition is achieved by simply computing the Euclidean distance between training and novel sheets. Similarly, Liu et al. extract two 2D sheets, namely the column projection (XT) and row projection (YT) of the binary silhouettes, and use them for recognition via K-nearest neighbor matching on the normalized correlation scores of training and novel sheets. He and Debrunner [18] compute a quantized vector of Hu moments from the person's binary silhouette at discrete gait poses, and use them for recognition via an HMM. The method of Kale et al. [28] is quite similar to this, except that they use the vector of silhouette widths (for each latitudes) instead of Hu moments. Certainly, the self-similarity plot used in the first method of the present paper is a mapping of the sequence of silhouettes to a 2D signal. However, while the SSP is quite robust to the segmentation noise in binary silhouettes, signals derived directly from binary silhouettes are typically sensitive to segmentation noise.

Another class of methods apply two levels of aggregation on the XYT data, and not one [16], [19], [23], [29]. They first map the XYT data of the walking person into one or more one-dimensional signals, then aggregate these into a feature vector by computing statistics of these signals (such as their first and second order moments). Lee and Grimson [29] fit ellipses to seven rectangular subdivisions of the silhouette then compute four statistics (first and second-order moments) for each ellipse, and hence obtain 28 one-dimensional signals from the entire silhouette sequence. They then use three different methods for mapping these 28 signals to obtain a more terse feature vector to use for classification.

Little and Boyd [16] use optical flow instead of binary silhouettes. They fit an ellipse to the dense optical flow of the person's motion, then compute thirteen scalars consisting of first and second order moments of this ellipse. Periodicity analysis is applied to the resulting thirteen 1D signals, and a 12-dimensional feature vector is computed consisting of the phase difference between one signal and all other twelve signals. Recognition is achieved via exemplar K-nearest neighbor classification in this 12-dimensional feature space. These features are both scale-invariant and time-shift invariant, so that no temporal scaling nor alignment is necessary.

Obviously, the advantage of the holistic approach lies in that it is correspondence-free, and hence simple to implement. Its main drawback is that the extracted features are inherently appearance-based, and are hence likely to be sensitive to any factors that alter the person's silhouette (whether color or binary), particularly camera viewpoint and clothing. Viewpoint

dependence can be remedied by estimating the viewpoint of the walking person and using view-based recognition. However, it is not obvious how or whether the clothing sensitivity problem could be solved.

B. Feature-based Approach

The feature-based approach recovers explicit features (or parameters) describing gait dynamics, such as stride dimensions and the kinematics of joint angles. Although human body measurements (i.e. absolute distances between certain landmarks, such as height, limb lengths, shoulder width, head circumference, etc.) are not descriptors of body movement, they are indeed determinants of that movement, and hence can also be considered as gait parameters.

Johnson and Bobick [22] compute body height, torso length, leg length and step length for identification. Using a priori knowledge about body structure at the double-support phase of walking (i.e. when the feet are maximally apart), they estimate these features as distances between fiducial points (viz. the midpoint and extrema) of the binary silhouette. Obviously, the accuracy of these measurements is very sensitive to segmentation noise in the silhouette, even if they are averaged over many frames.

In [42], Davis uses a similar approach to compute the stride length and cadence, though he relies on reflective markers to track 3D trajectories head and ankle. With measurements obtained from 12 people, he is able to train a linear perceptron to discriminate the gaits of adults and children (3-5 years old) to within 93% accuracy.

Benabdelkader et al. describe a more robust method to compute stride dimensions, which exploits not only the periodicity of walking but also the fact that people walk in contiguous steps [44]. In related work [26], they further estimate the height variation of a walking person by fitting it to a sinusoidal model, and use the two model parameters along with the stride dimensions for identification.

The kinematics of a sufficient number of body landmarks can potentially provide a much richer, and perhaps unique, description of gait. Bissacco et al. [27] fit the trajectories of 3D joint positions and joint angles to a discrete-time continuous-state dynamical system. They use the space spanned by the parameters of this model for recognizing different gaits. Tsai et al. [41] use one cycle of the spatio-temporal curvature function of 3D trajectories of certain points on the body for identification.

The major strength of this approach lies in that, unlike its holistic counterpart, it uses classification features that are known to be directly pertinent to gait dynamics. Another advantage is that it is in principle view-invariant since it uses 3D quantities for classification. However, its measurement accuracy degrades for certain viewpoints as well as at low resolutions. Obviously, accurate measurement of most of these gait parameters requires not only accurate camera calibration but also accurate detection and tracking of anatomical landmarks in the image sequence. The feasibility of this approach is currently very limited mainly due to the difficulty of automatic detection and tracking in realistic (low-resolution) video. For example in [41] [42] and [27], the authors use 3D motion capture data or semi-manually tracked features in order to avoid the automatic detection and tracking problem.

III. Method

This method is based on the *self-similarity plot* (SSP), defined as the matrix of cross-correlation between each pair of images of the person in the sequence. By properly aligning and scaling the

SSP (to account for differences in gait frequency and phase), we extract normalized feature vectors which we subsequently use as input to a standard pattern classifier for recognition.

The method is deemed holistic because the SSP is a direct transformation of the spatiotemporal volume (XYT) of the walking person. Computation of the SSP is correspondence-free and is robust to segmentation and tracking errors. Intuitively, the SSP encodes both the static (first-order) properties as well as temporal variations of body shape during the walking, and hence can be regarded as a 2D signature of gait. Furthermore, we contend that the SSP is a projection of planar gait dynamics, provided the person is sufficiently far from the camera [47].

For the sake of clarity, we present the method in terms of the structure of a general pattern classifier [48], [49], as shown in Figure 1. The raw input, an image sequence of a walking person, is processed sequentially by three main modules: (1) a *pre-processing* module that segments and tracks the moving person in each frame, (2) a *feature measurement* module which computes the self-similarity plot and extracts normalized features from it, and (3) a *pattern classification* module which determines the identity of the walking person based on the given features.

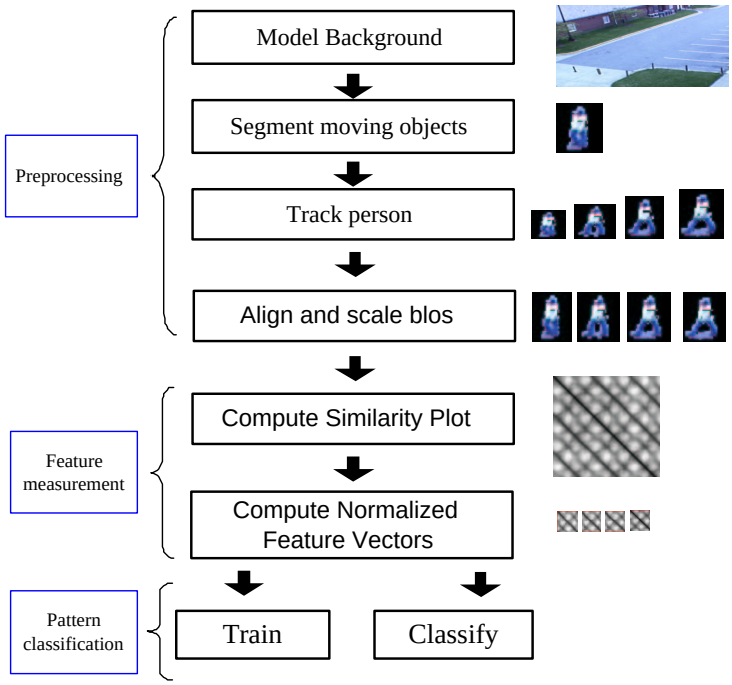


Fig. 1. Overview of Method.

A. Pre-processing

A.1 Segmentation and Tracking

Given a sequence of images obtained from a static camera, we detect and track the moving person, and compute the corresponding sequence of motion regions (or blobs) in each frame. Motion segmentation is achieved via a non-parametric background modeling/subtraction technique that is quite robust to lighting changes, camera jitter and to the presence of shadows [50]. Once detected, the person is tracked in subsequent frames via simple spatial coherence, namely based on the overlap of blob bounding boxes in any two consecutive frames [51]. The issue of determining whether a foreground blob indeed corresponds to a moving person is addressed in

the feature measurement module[1]. Specifically, we use the cadence-based technique described in [35] which simply verifies whether the computed cadence is within the normal range of human walking (roughly $[80, 145]$ steps/min).

A.2 Template Extraction

Once a person has been tracked for N consecutive frames, a sequence of N corresponding templates is created as follows. Given the person's blob in each frame, we extract the (rectangular) region[2] enclosed within its bounding box either from (1) the original color/greyscale image, (2) the foreground image, or (3) the binary image, as shown in Figure 2. Clearly, there are competing tradeoffs to using either type of template in measuring image similarity (when computing the SSP). The first is more robust to segmentation errors. The third is more robust to clothing and background variations. The second is simply a hybrid of these two; it is robust to background variations but sensitive to segmentation errors and clothing variations.

Fig. 2. Template types from left-to-right: original image, foreground template and binary template, of a walking person.

B. Feature Measurement

Given the sequence of N templates cropped from the N image frames of a walking person, we compute a NxN matrix of their pairwise correlations, denoted the self-similarity plot (SSP). In the following sections, we explain the different steps for computing and normalizing the SSP to finally obtain features that we can use for recognition.

B.1 Template Scaling

Since the template size varies according to camera viewpoint and depth[3], as shown in Figure 3, we scale the templates prior to computing the SSP as explained below. It is important to note, however, that this achieves scale-invariance only in the case of small depth changes. Large depth changes introduce other variations that cannot be normalized by mere scaling. For example, inherently less detail of the gait dynamics is captured as the depth increases. Also, perspective effects are more prominent at small depths.

[1]The only reason this is *not* done in the current module is for the sake of modularity, since cadence is computed in the second module.

[2]The cropped region also includes an empty 10-pixel border in order to allow for shifting when we later compute the cross-correlation of template pairs.

[3]It also varies with image resolution, which we assume is constant.

The apparent size of a walking person varies at the frequency of gait, due to the pendular-like oscillatory motion of the legs and arms, and consequently the width and height of a person's image also vary at the fundamental frequency of walking. Specifically, let $w(n)$ and $h(n)$ be the width and height of the n-th image (template) of the person. According to gait analysis literature [6], $w(n)$ and $h(n)$ can be approximated as sinosoidal functions:

$$\begin{aligned} w(n) &= m_w(n) + A_w \sin \omega n + \phi \\ h(n) &= m_h(n) + A_h \sin \omega n + \phi \end{aligned}$$

where ω is the frequency of gait (in radians per frame) and ϕ is the phase of gait (in radians). Note that $m_w(n)$ is the mean width and A_w is the amplitude of oscillation (around this mean). The same can be said about $m_h(n)$ and A_h, respectively, for height. Furthermore, in fronto-parallel walking, $m_w(n)$ and $m_h(n)$ are almost constant, while in non-fronto-parallel walking, and due to the changing camera depth, they increase/decrease approximately linearly (i.e. in a linear trend): $m_w(n) \approxeq \alpha_w n + \beta_w$ and $m_h(n) \approxeq \alpha_h n + \beta_h$. Figure 3 illustrates these two different cases.

Therefore, in order to account for template size variation caused by camera depth changes (during non-fronto parallel walking), we first de-trend them:

$$\begin{aligned} \hat{w}(n) &= w(n) - \alpha_w n = \beta_w + A_w \sin \omega n + \phi \\ \hat{h}(n) &= h(n) - \alpha_h n = \beta_h + A_h \sin \omega n + \phi \end{aligned}$$

so that the templates now have equal mean width and height. Note, however, that we need $\frac{\hat{w}(n)}{w(n)} = \frac{\hat{h}(n)}{h(n)}$ for all n, i.e. $\frac{\alpha_w}{\alpha_h} = \frac{w(n)}{h(n)}$, so that each template can be uniformly scaled along its width and height. In other words, we need the width-to-height aspect ratio to remain constant throughout the sequence. This is a valid assumption since the person is sufficiently far from the camera, and barring abrupt/sharp changes in person's pose with respect to the camera.

Finally, the templates are scaled one more time so that their mean height is equal to some given constant H_0 (we typically use $H_0 = 50$ pixels):

$$\tilde{h}(n) = \hat{h}(n) \cdot \frac{H_0}{\beta_h} = H_0 + \tilde{A}_h \sin \omega n + \phi \tag{1}$$

B.2 Computing the Self-similarity Plot

Let I_i be the ith scaled template with size $\tilde{w}_i \mathrm{x} \tilde{h}_i$ (in pixels). The corresponding self-similarity plot $S(i,j)$ is computed as the absolute correlation[4] of each pair of templates I_i and I_j, minimized over a small search radius r, namely:

$$S(i,j) = \min_{|dx|<r, |dy|<r} \sum_{|x| \le \frac{W}{2}} \sum_{|y| \le \frac{H}{2}} |I_j(x+dx, y+dy) - I_i(x,y)|. \tag{2}$$

where $W = min(\tilde{w}_i, \tilde{w}_j - 2r)$ and $H = min(\tilde{h}_i, \tilde{h}_j - 2r)$ so that the summation does not go out of bounds. Although ideally S should be symmetric, it typically is not, unless $r = 0$.

[4]We chose absolute correlation for its simplicity. Other similarity measures include normalized cross-correlation, the ratio of overlapping foreground pixels, Hausdorff distance, etc.

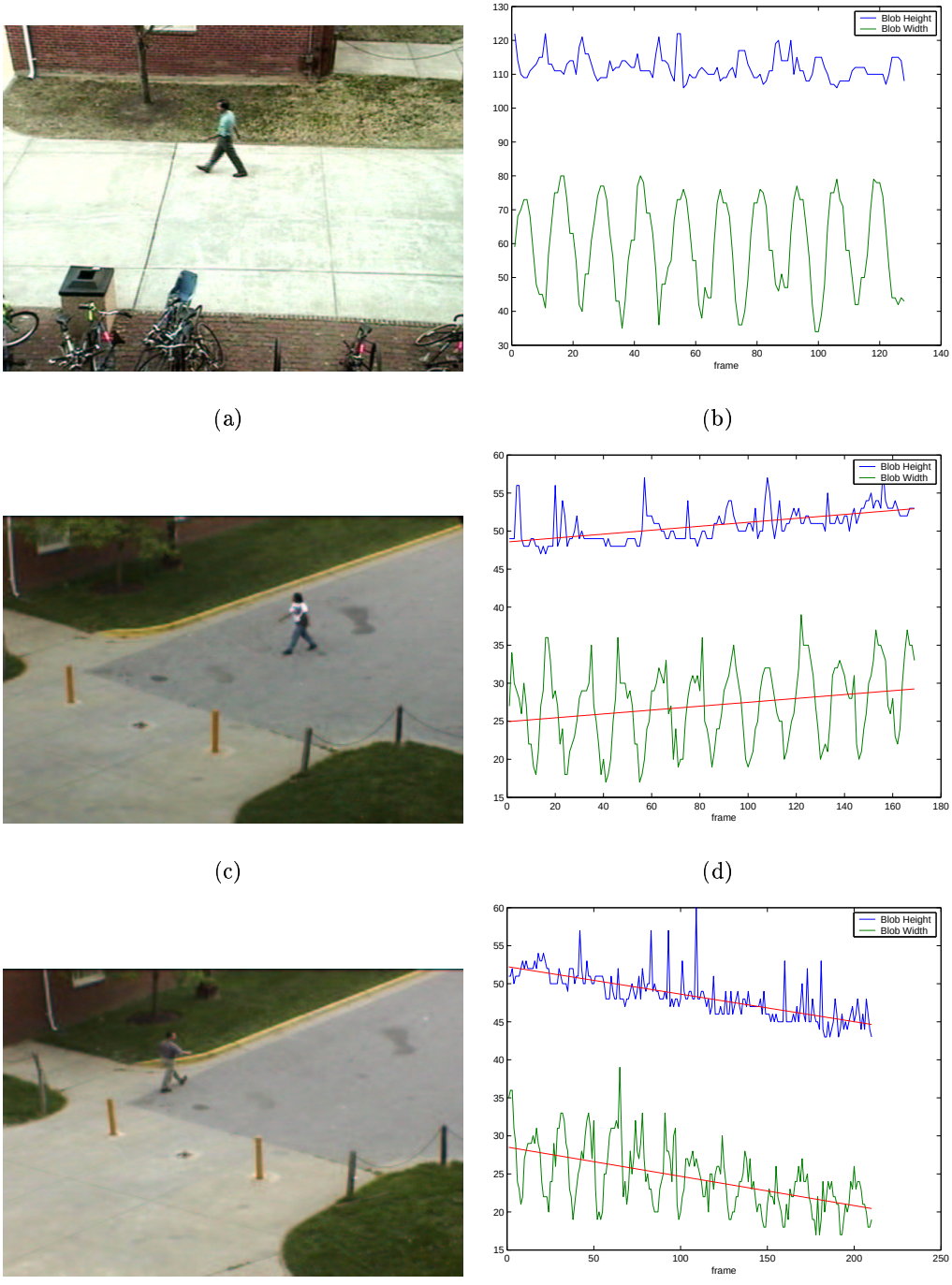
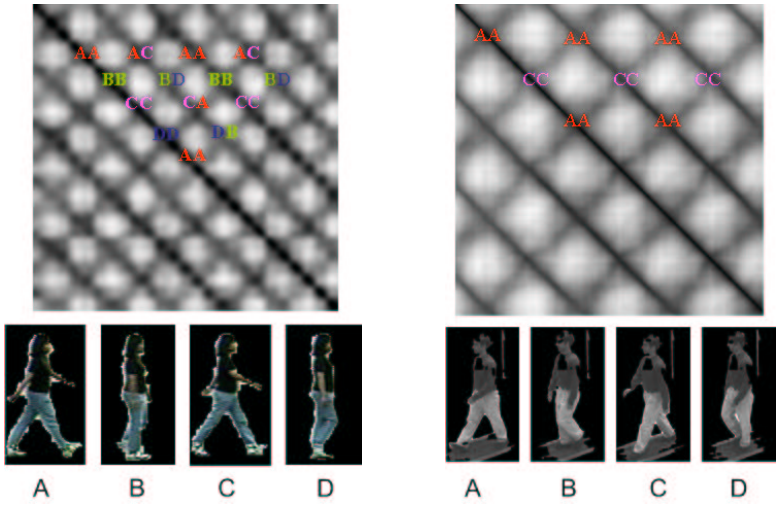


Fig. 3. Template dimensions in pixels for: (a,b) a fronto-parallel sequence, (c,d,e,f) two non-fronto-parallel sequences (bottom two rows). The width and height increase when the person walks closer to the camera (middle row), and decrease as the person moves away from the camera (bottom row). The red lines correspond to the linear trend in both these cases.

Figure 4 highlights some of the properties of S for fronto-parallel and non-fronto-parallel walking sequences. The diagonals are due to the periodicity of gait, while the cross-diagonals are due to the temporal mirror symmetry of the gait cycle [47]. The intersections of these diagonals, i.e. the local minima of S, correspond to key poses of the gait cycle, the mid-stance (B and D) and double-support (A and C) poses. Thus S encodes both the frequency and phase of the gait cycle. Some of these intersections disappear for non-fronto-parallel sequences (BD, BB and DD) because gait does not appear bi-laterally symmetric.



(a) (b)

Fig. 4. The self-similarity plots for (a) a fronto-parallel sequence, and (b) a non-fronto-parallel sequence, computed here using foreground templates. Similarity values are linearly scaled to the grayscale intensity range [0,255] for visualization. The local minima of each SSP correspond to combinations of key poses of gait (labelled A, B, C and D).

B.3 Normalizing the Self-similarity Plot

Since we are interested in using the SSP for recognition, we need to be able to compare the SSP's of two different walking sequences. Furthermore, gait consists of repeated steps, and so it only makes sense to compare two SSP's that contain an *equal* number of walking cycles and start at the *same* phase (i.e. body pose). In other words, we need to normalize the SSP for differences in sequence length and starting phase. There are several ways to achieve this. In a previous work, we used a sub-matrix of the SSP that starts at the first occurrence of the double-support pose[5] in the sequence and spans three gait cycles (i.e. six steps) [52].

A different approach that proves to be better for recognition [25] uses the so-called Self-Similarity Units (SSU). Each SSU is a sub-matrix of the SSP that starts at the double-support phase and spans one gait cycle. The SSP can then be viewed as a tiling of (contiguous) SSU's, and a different tiling can be obtained for any particular starting phase. We use all SSU's corresponding to the left and right double-support poses for gait recognition. However, because the SSP is (approximately) symmetric and for computational efficiency, we only use the SSU's of the top half, as shown in Figure 5. We can easily show that for a sequence containing K gait cycles,

[5]The double-support phase of the gait cycle corresponds to when the feet are maximally apart. The left double-support pose is when the left leg is leading and the right double-support pose is when the right leg is leading.

there are $2\frac{K(K+1)}{2} = K(K+1)$ SSU's.

Finally, because the size of each SSU is defined both by the duration of a gait cycle and the frame rate (namely $P = T \cdot F_s$ frames, where T is the average gait cycle length in seconds and F_s is the frame rate), we scale all SSU's to some uniform size of mxm in order to be able to compare them.

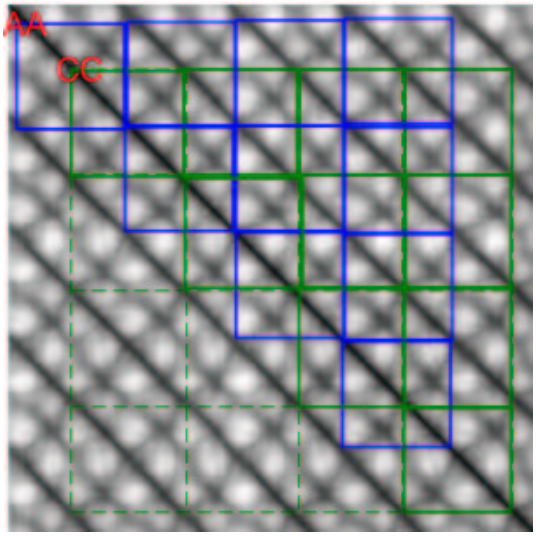


Fig. 5. Extracting self-similarity units from the similarity plot. Blue and green SSU's start at pose A and C, respectively.

B.4 Computing the Frequency and Phase of Gait

Obviously, we need to compute the frequency and phase of gait in order to normalize the SSP and obtain the SSU's. Several methods in the vision literature have addressed this problem, typically via periodicity analysis of some feature of body shape or texture [53], [54], [12]. In fact, most existing gait recognition methods involve some type of frequency/phase normalization, and hence devise some method for computing the frequency and phase of gait.

In this paper, we compute gait frequency and phase via analysis of the SSP, which indeed encodes the frequency and phase of walking, as mentioned in Section III-B.2. We found this to be more robust than using say the width or height of the silhouette, as we have done in the past [52]. For the frequency, we apply the autocorrelation method on the SSP as was done in [12]. This method is known to be more robust to non-white noise and non-linear amplitude modulations than Fourier analysis. It first smooths the autocorrelation matrix of the SSP, computes its peaks, then finds the best-fitting regular 2D lattice for these peaks. The period is then obtained as the width of this best-fitting lattice.

The phase is computed by locating the local minima of the SSP that correspond to the A and C poses (defined in Section III-B.2). However, not *all* local minima correspond to these two poses, since in near-fronto-parallel sequences combinations of the B and D poses also form at local minima. Fortunately, the two types of local minima can be distinguished by the fact that those corresponding to A and C poses are 'flatter' than those corresponding to B and D poses. However, we are still only able to resolve the phase of gait up to half a period, since we have no way of distinguishing the A and C poses from one another. As a result, the SSU's corresponding to both the A and C poses (shown in Figure 5) are all used for gait recognition.

C. Pattern Classification

We formulate the problem as one of supervised pattern classification. Given a set of labelled sequences (the *gallery*), we determine the person (class) corresponding to a novel sequence (the *probe*) using a pattern classifier that takes their respective SSU's as the input patterns. For classification, we use the following variation of the K-nearest neighbor (KNN) rule [55]. Let G be the set of all SSU's corresponding to the sequences in the gallery, and suppose l SSU's are extracted from the probe sequence. For each probe SSU, we find the K closest neighbors in G, based on some distance (or dissimilarity) metric. We then assign the probe to the majority label of the total $l \cdot K$ neighbors. In the sequel, we discuss two different approaches for measuring the distance between two SSU's: template matching and statistical pattern classification [48], [55], as shown in Figure 6.

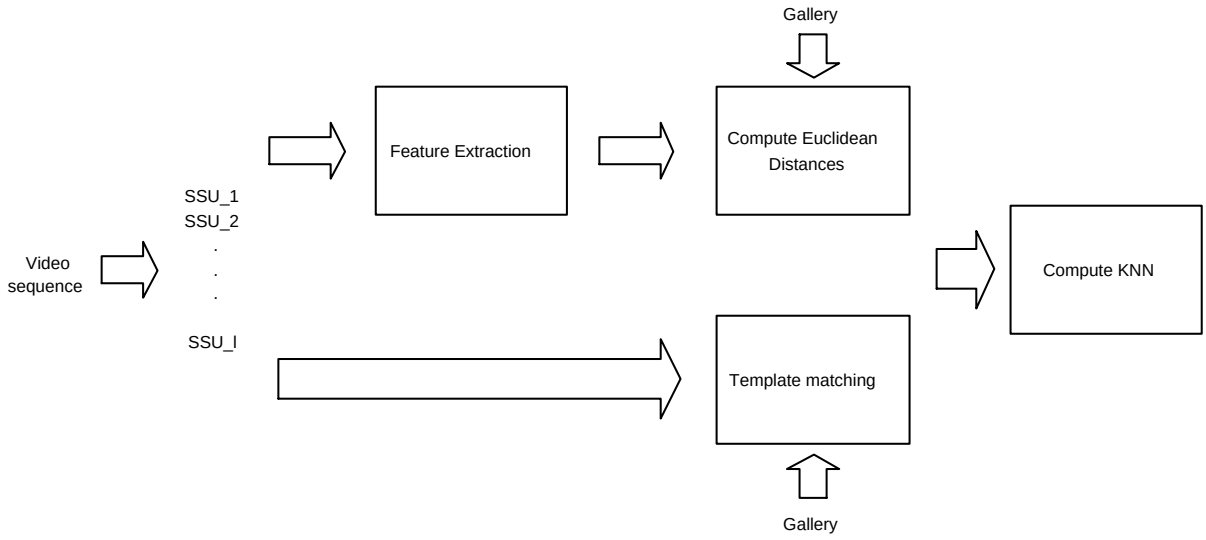


Fig. 6. The pattern classifier used to recognize a novel video sequence (the probe) based on a set of labelled sequences (the gallery). The input patterns are the SSU's derived from each sequence and classification is based on the K-nearest neighbor (KNN) rule. We consider two different approaches to computing the inter-pattern distances: template matching and Euclidean distances in a reduced subspace.

C.1 Template Matching

Because the SSU is a mxm 2D template, perhaps the simplest distance metric between two SSU's is their maximum cross-correlation computed over a small range of 2D shifts (we typically use the range $[-5, 5]$). The advantage of this approach is that it explicitly compensates for small phase alignment errors. Its disadvantage is that it is computationally very demanding.

C.2 Statistical Pattern Classification

Here, each SSU is represented as a p-dimensional vector, $p = m^2$, by concatenating its m rows. The distance between two patterns is then simply computed as their Euclidean distance in this space. However, when p is large, it is desirable to first reduce the dimensionality of the vector space for the sake of computational efficiency, as well as to circumvent the curse of dimensionality phenomenon [48], [55], [49].

Dimensionality reduction, also called feature extraction, maps the vectors to a q-dimensional

space, with $q \ll p$. We consider three linear feature extraction techniques for this problem: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and a so-called subspace-LDA that combines the latter two techniques by applying LDA on a subspace spanned by the first few principal components. See [56], [57], [58], [59], [60], [61] for examples of the application of these methods in face recognition.

Each method defines a linear transformation W that maps a p-dimensional vector u in the original feature space onto a q-dimensional vector $\zeta = (\zeta_1, .., \zeta_q)$ such that $\zeta = W^T u$. Note that $(\zeta_1, .., \zeta_q)$ can also be viewed as the coordinates of u in this q-dimensional subspace.

The matrix pxq matrix W is determined from a given training set of vectors by optimizing some objective criterion. The choice of q seems to be domain-dependent and we have not as yet devised a method to automatically select it. Instead we simply choose the value that achieves best classification rate for the given training and test data sets.

Choosing between PCA, LDA and subspace-LDA is also domain-dependent. It depends on the relative magnitudes of the within-class scatter and the between-class scatter, as well as the size of the training set. Furthermore, one design issue common to all three approaches is the choice of the subspace dimensionality.

IV. Experiments and Results

We evaluate the performance of the method on four independently acquired data sets of varying degrees of difficulty. Our goal is to quantify the effect of the following factors on performance:
1. Natural individual variability, caused by various physical and psychological factors (clothing, footwear, cadence, mood, fatigue, etc.). This within-person variation is introduced by using multiple samples of each person's walking taken at different times and/or over different days. However, sequences taken on different days may also contain extra unwanted variation (i.e. that is irrelevant to gait recognition) which makes recognition more difficult. For example, background subtraction noise (in the segmented silhouettes) is often a function of the background scene and lighting conditions, both of which may change over different imaging sessions. Change in clothing also introduces unwanted variation since our method is appearance-based.
2. Photometric parameters, viz. camera viewpoint, camera depth, and frame sampling rate. Since our method is holistic and hence appearance-based, we expect its performance to be sensitive to any appearance-altering factors (which certainly includes camera parameters).
3. Algorithm design parameters, viz. the image similarity measure (correlation of binary silhouettes and correlation of foreground silhouettes), the pattern representation approach (PCA, LDA, s-LDA and template matching), and the KNN classifier parameter ($K = 1, 3, 5$).
We use the split-sample (hold-out) or the leave-one-out cross-validation methods to estimate the classification error rate for each data set [62], [63], [55].

A. Dataset 1

This data set is the same used by Little and Boyd in [16]. It consists of 42 image sequences with six different subjects (4 males and two females), 7 sequences of each, taken from a static camera at 30 fps and 320x240 resolution. The subjects walked a fixed path against a uniform background. Thus the only source of variation in this data set (aside from random measurement noise) is the individuals' own walking variability across different samples.

Figure 7 shows all seven subjects overlaid on the background image. The results are shown in Table I. Note that LDA is not used for this data set because the number of training samples does not satisfy the condition in (**??**). Obviously BC gives slightly better results than FC, and

that subspace-LDA also slightly outperformed PCA. However, there is a significant improvement when using feature extraction (PCA and s-LDA) over template matching (TM).

Fig. 7. The six subjects for dataset 1, shown overlaid on the background image.

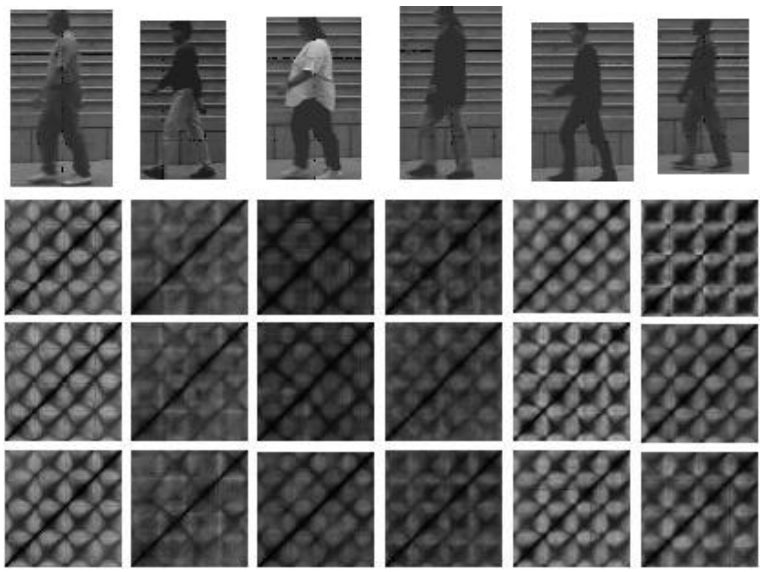


Fig. 8. Three of the self-similarity plots for each person in dataset 1.

B. Dataset 2

The second data set contains fronto-parallel sequences of 44 different subjects (10 females and 34 males), taken in an outdoor environment from two different cameras simultaneously, as shown in Figure 9. The two cameras are both fronto-parallel but located at different depths (approximately 20 ft and 70 ft) with respect to the walking plane. Each subject walked in two different sessions a fixed straight path, back and forth, at their natural pace. The sequences were captured at 20 fps and a full-color resolution of 644x484.

Table II summarizes the classification results of six different hold-out experiments. Absolute correlation of binary silhouettes (BC) is used in all experiments. The classification performance is better for the far camera (first row) that for the near camera (second row). We suspect this may be because the templates obtained in the far camera are smaller and hence perhaps less affected by segmentation noise. Furthermore, the performance degrades significantly when the train and

TABLE I

CLASSIFICATION RATES FOR THE FIRST DATA SET FOR DIFFERENT PATTERN REPRESENTATION APPROACHES (PCA, S-LDA AND TM), AND DIFFERENT VALUES OF THE KNN CLASSIFIER PARAMETER K.

	Leave-one-out					
	BC			*FC*		
K	PCA	S-LDA	TM	PCA	S-LDA	TM
1	93	98	92	95	95	90
3	95	98	95	95	93	93
5	95	98	95	95	93	93

test sets are from different cameras (third and fourth rows). Thus our method is not invariant to large changes in camera depth. This confirms our expectation that depth/POT invariance of the scaled templates is only good up to small variations in camera depth (Section III-B.1).

(a) (b)

Fig. 9. Second outdoor dataset. Sample frames from (a) the *near* camera and (b) the *far* camera.

C. Dataset 3

In order to evaluate the performance of the method across large changes in camera viewpoint, we used the Keck multi-perspective lab [64] to capture sequences of people walking on a treadmill from 8 different cameras at a time, as illustrated in Figure 10. The cameras are placed at the same height around half-a-circle so that they have the same tilt angle and different pan angles. The latter span a range of about 135 deg of the viewing sphere, though not uniformly. The data set contains 12 people (3 females and 9 males) and about 5 sequences per person per view on average, taken mostly on different days for each person. The sequences were captured at a frame rate of 60 fps and a resolution of 644x488 greyscale images.

Like in general object recognition problems, there are two main approaches to gait recognition under variable viewing conditions: a view-based approach and a parametric approach. In the view-based approach, a classifier is trained separately for each viewpoint, i.e. there are as many classifiers as there are camera viewpoints. A novel sequence needs to first have its viewpoint determined so that the corresponding classifier is applied. The parametric approach, on the other hand, trains a single classifier using data from all viewpoints.

TABLE II

CLASSIFICATION PERFORMANCE ON THE SECOND DATA SET USING HOLD-OUT TECHNIQUE WITH SIX DIFFERENT TRAINING AND TESTING SUBSETS.

		PCA		LDA		s-LDA		TM
Train set	Test set	K=1	K=3	K=1	K=3	K=1	K=3	K=1
Far Camera	Far Camera	49	49	63	70	65	67	59
Near Camera	Near Camera	53	52	41	46	49	52	55
Far camera	Near camera	10	10	22	23	24	25	21
Near camera	Far camera	17	23	24	22	23	23	25
Both cameras	Both cameras	52	50	52	52	54	56	35

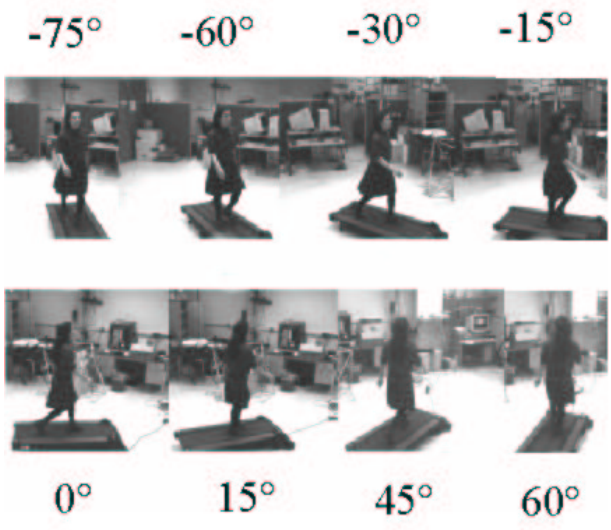
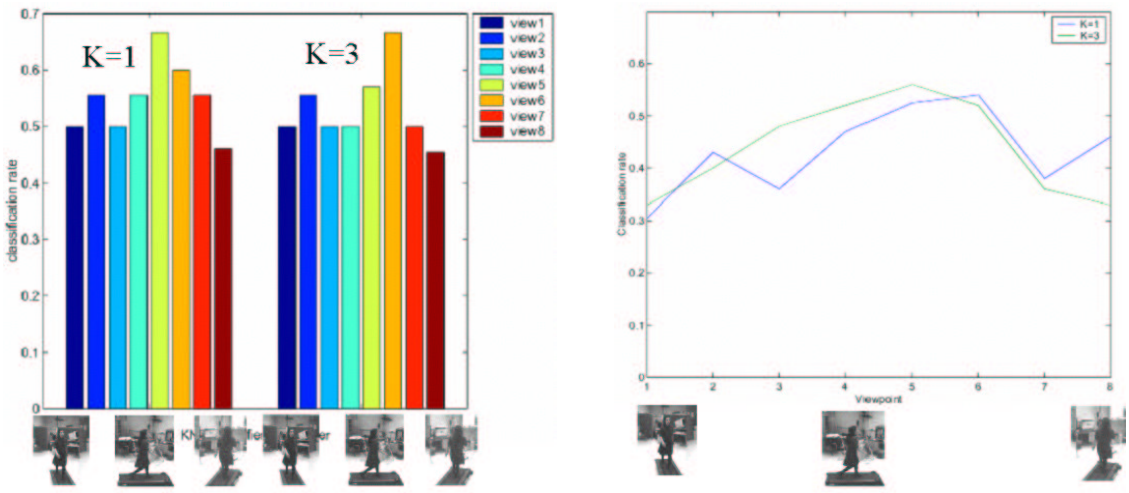


Fig. 10. Eight camera viewpoints of the sequences in second test data set.

Both these approaches are applied to the dataset and the results are shown in Figure 11. We use absolute correlation of binary silhouettes for image similarity, and the hold-out cross-validation technique to estimate the classification rate, whereby we train on data from six days and test on data from the 7th day. This is repeated 7 times, and the classification rate is computed as the average over the seven iterations. Clearly, the performance is best for near-fronto-parallel views (4-6). An intuitive explanation for this is the following. Most of the dynamics of walking takes place in the sagittal plane, which is the plane containing both legs. Hence in a non-fronto-parallel viewpoint, where the sagittal plane is almost orthogonal to the image plane, much less of the appearance variation caused by gait dynamics is captured, and which may be insufficient for recognition. Furthermore, view-based approach gives overall better results than the parametric approach.

D. Dataset 4

Recall that we scale the SSU's to a fixed size (mxm pixels), which is equivalent to normalizing the gait frequency to a fixed value (i.e. temporal scaling). However, because gait dynamics are inherently a function of cadence, we expect that the SSU's corresponding to significantly different



(a) (b)

Fig. 11. Classification performance for Dataset 3: (a) View-based approach. (b) Parametric approach.

cadences to be *qualitatively* different (even if they are pre-normalized to the same frequency). We tested this expectation using a portion of CMU's MoBo dataset [65], consisting of indoor sequences of 25 people walking on a treadmill and captured from 3 different views, as shown in Figure 12. Furthermore, each person walked at two different speeds: a slow pace (2.06 miles/hr) and a moderate pace (2.82 miles/hr). Thus we used a total of 150 sequences for this experiment (i.e. 2 sequences per person per view). The sequences are all captured on the same day and against the same background.

(a) (b) (c)

Fig. 12. The three views in Dataset 4.

We then setup experiments via the hold-out cross-validation technique in which the train and test sets correspond to different combinations of speeds. This was done for each view separately, however (i.e. view-based classification). Table III shows the classification results. As expected, the performance degrades significantly when the train and test data for the classifier correspond to different speeds.

TABLE III

PERFORMANCE ON THE FIFTH DATASET: CLASSIFICATION RATES USING HOLD-OUT TECHNIQUE WITH FOUR DIFFERENT TRAINING AND TESTING SUBSETS.

	Slow/Slow	Fast/Fast	Slow/Fast	Fast/Slow
View 1	100	100	54	32
View 2	100	96	26	16
View 3	96	100	43	33

V. Conclusions

We described a novel holistic gait recognition approach that uses image self-similarity as the basic feature for classification. The method is correspondence-free, works well with low-resolution video, and is robust to to variation in clothing, lighting, and to segmentation errors. A recognition rate of 98% is achieved for a fronto-parallel data set of 6 people, and 70% for a fronto-parallel data set of 54 people.

Although the method is inherently appearance-based, and hence view-dependent, this is circumvented via view-based recognition. Using a data set of 12 people captured from eight viewpoints, recognition rates decreases from about 65% for near fronto-parallel viewpoints to about 47% for near frontal viewpoints. Performance also degrades when camera depth and cadence are significantly changed.

We are working to combine the gait features of this method with geometric gait features that can be robustly computed from video, such as cadence, stride length and stature. We also plan to study the use of these features for other recognition tasks, such as gender classification and gait asymmetry detection (such as caused by a limp).

VI. Acknowledgments

This paper was written under the support of DARPA's HumanID at a Distance Project.

References

[1] A. Jain, *Biometrics: Personal Identification in Networked Society*, Kluwer Academic, Norwell, 1999.
[2] D. D. Zhang, *Automated Biometrics: Technologies and Systems*, Kluwer Academic Publishers, 2000.
[3] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception and Psychophysics*, vol. 14, no. 2, 1973.
[4] J.E. Cutting and L.T. Kozlowski, "Recognizing friends by their walk: Gait perception without familiarity cues," *Bulletin Psychonomic Soc.*, vol. 9, no. 5, pp. 353–356, 1977.
[5] J. Cutting C. Barclay and L. Kozlowski, "Temporal and spatial factors in gait perception that influence gender recognition," *Perception and Psychophysics*, vol. 23, no. 2, pp. 145–152, 1978.
[6] M. Murray, "Gait as a total pattern of movement," *American Journal of Physical Medicine*, vol. 46, no. 1, 1967.
[7] H. J. Ralston V. Inman and F. Todd, *Human Walking*, Williams and Wilkins, 1981.
[8] D. Winter, *The Biomechanics and Motor Control of Human Gait*, Univesity of Waterloo Press, 1987.
[9] J. Perry, *Gait Analysis: Normal and Pathological Function*, SLACK Inc., 1992.
[10] J. Rose and J. G. Gamble, *Human Walking*, Williams and Wilkins, 1994.
[11] C. Cedras and M. Shah, "A survey of motion analysis from moving light displays," in *Proceedings of the Computer Vision and Pattern Recognition*, 1994.
[12] R. G. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 2, 2000.
[13] S. Niyogi and E. Adelson, "Analyzing gait with spatiotemporal surfaces," in *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, 1994.
[14] S. Niyogi and E. Adelson, "Analyzing and recognizing walking figures in XYT," in *Proceedings of the Computer Vision and Pattern Recognition*, 1994.
[15] H. Murase and R. Sakai, "Moving object recognition in eigenspace representation: gait analysis and lip reading," *Pattern Recognition Letters*, vol. 17, 1996.
[16] J. Little and J. Boyd, "Recognizing people by their gait: the shape of motion," *Videre*, vol. 1, no. 2, 1998.

[17] P. S. Huang, C. J. Harris, and M. S. Nixon, "Comparing different template features for recognizing people by their gait," in *BMVC*, 1998.
[18] Q. He and C. Debrunner, "Individual recognition from periodic activity using hidden markov models," in *IEEE Workshop on Human Motion*, 2000.
[19] M. S. Nixon James B. Hayfron-Acquah and John N. Carter, "Recognising human and animal movement by symmetry," in *Audio- and Video-based Biometric Person Authentication*, 2001.
[20] M.S. Nixon D. Cunado and J.N. Carter, "Gait extraction and description by evidence gathering," in *Audio- and Video-based Biometric Person Authentication*, 1999.
[21] M. S. Nixon C. Yam and J. N. Carter, "Extended model-based automatic gait recognition of walking and running," in *Audio- and Video-based Biometric Person Authentication*, 2001.
[22] A. Johnson and A. Bobick, "Gait recognition using static activity-specific parameters," in *Proceedings of the Computer Vision and Pattern Recognition*, 2001.
[23] D. Zlatnik P. C. Cattin and R. Borer, "Biometric system using human gait," in *Mechatronics and Machine Vision in Practice (M2VIP)*, 2001.
[24] J. Boyd, "Video phase-locked loops in gait recognition," in *Proceedings of the Computer Vision and Pattern Recognition*, 2001.
[25] R. G. Cutler C. BenAbdelkader and L. S. Davis, "Motion-based recognition of people in eigengait space," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.
[26] R. G. Cutler C. BenAbdelkader and L. S. Davis, "View-invariant estimation of height and stride for gait recognition," in *Post-ECCV Workshop on Biometric Authentication*, 2002.
[27] A. Bissacco, Y. Ma A. Chiuso, and S. Soatto, "Recognition of human gaits," in *Proceedings of the Computer Vision and Pattern Recognition*, 2001.
[28] N. Cuntoor V. Kruger A. Kale, A.N. Rajagopalan and R. Chellapa, "Human identification using gait," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.
[29] L. Lee and W.E.L. Grimson, "Gait apperance for recognition," in *Post-ECCV Workshop on Biometric Authentication*, 2002.
[30] R. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.
[31] Y. Liu, R. Collins, and Y. Tsin, "Gait sequence anaysis using frieze patterns," in *European Conference on Computer Vision*, 2002.
[32] P.J. Philips S. Sarkar I. Robledo, P. Grother and K. Bowyer, "Baseline results for the challenge problem of human id using gait analysis," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.
[33] Q. Cai and J. K. Aggarwal, "Human motion analysis: a review," in *Proc. of IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects*, 1997.
[34] Dariu Gavrila, "The visual analysis of human movement: a survey," *Computer Vision and Image Understanding*, vol. 73, no. 1, 1999.
[35] S. Yasutomi and H. Mori, "A method for discriminating pedestrians based on rythm," in *IEEE/RSG Intl Conf. on Intelligent Robots and Systems*, 1994.
[36] X. Feng Y. Song and P. Perona, "Towards detection of human motion," in *Proceedings of the Computer Vision and Pattern Recognition*, 2000.
[37] L. W. Campbell and A. Bobick, "Recognition of human body motion using phase space constraints," in *International Conference on Computer Vision*, 1995.
[38] J. W. Davis, "Appearance-based motion recognition of human actions," M.S. thesis, Media Arts and Sciences, MIT, 1996.
[39] J. Psl D. Meyer and H. Niemann, "Gait classification with hmms for trajectories of body parts extracted by mixture densities," in *British Machine Vision Conference*, 1998.
[40] N. Cuntoor A. Kale and R. Chellapa, "A framework for activity based human recognition," in *International Conference on Acoustics Speech and Signal Processing*, 2002.
[41] K. Keiter P. Tsai, M. Shah and T. Kasparis, "Cyclic motion detection for motion–based recognition," *Pattern Recognition*, vol. 27, no. 12, 1994.
[42] J. W. Davis, "Visual categorization of children and adult walking styles," in *Audio- and Video-based Biometric Person Authentication*, 2001.
[43] M. S. Nixon C. Yam and J. N. Carter, "Gait recognition by walking and running: a model-based approach," in *Asian Conference on Computer Vision*, 2002.
[44] R. G. Cutler C. BenAbdelkader and L. S. Davis, "Stride and cadence as a biometric in automatic person identification and verification," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.
[45] K. Deffenbacher A.J. O'Toole, H. Abdi and D. Valentin, *A Percpetual Learning Theory of the Information in Faces*, chapter 8, pp. 159–182, London: Routledge, 1995.
[46] W. Zhao, "Face recognition: A literature survey," Tech. Rep. CAR-TR-948, UMCP, 2000.
[47] R. G. Cutler, *On the Detection, Analysis, and Applications of Oscillatory Motions in Video Sequences*, Ph.D. thesis, Phd Dissertation, University of Maryland, College Park, 2000.
[48] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, New York Academic Press, 1990.
[49] P. Hart R. Duda and D. Stork, *Pattern Classification*, John Wiley and Sons, 2001.
[50] D. Harwood A. Elgammal and L. S. Davis, "Non-parametric model for background subtraction," in *ICCV*, 2000.

[51] D. Harwood I. Haritaoglu and L. S. Davis, "W4s: A real-time system for detecting and tracking people in 21/2 d," in *European Conference on Computer Vision*, 1998.
[52] R. G. Cutler C. BenAbdelkader and L. S. Davis, "Eigengait: Motion-based recognition of people using image self-similarity," in *Audio- and Video-based Biometric Person Authentication*, 2001.
[53] R. Polana and R. Nelson, "Detection and recognition of periodic, non-rigid motion," *International Journal of Computer Vision*, vol. 23, no. 3, 1997.
[54] D. Harwood I. Haritaoglu, R. G. Cutler and L. S. Davis, "Backpack: Detection of people carrying objects using silhouettes," *Computer Vision and Image Understanding*, vol. 6, no. 3, 2001.
[55] R. P. W. Duin A. K. Jain and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, January 2000.
[56] I. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Journal of Optical Society of America A*, vol. 4, no. 3, pp. 519–524, March 1987.
[57] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, 1991.
[58] D.L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 3, pp. 831–836, August 1996.
[59] J. Hespanha P. Belhumeur and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, July 1997.
[60] R. Chellappa W. Zhao and P.J. Phillips, "Subspace linear discriminant analysis for face recognition," Tech. Rep. CAR-TR-914, UMCP, 1999.
[61] A.M Martinez and A.C. Kak, "Pca versus lda," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, February 2001.
[62] B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, 1996.
[63] S.M. Weiss and C.A. Kulikowski, *Computer Systems that Learn*, Morgan Kaufman, 1991.
[64] E. Borovikov, R. G. Cutler, T. Horprasert, and L. S. Davis, "Multi-perspective analysis of human actions," in *Third International Workshop on Cooperative Distributed Vision*, 1999.
[65] R. Gross and J. Shi, "The cmu motion of body (mobo) database," Tech. Rep. CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, 2001.