CHAPTER 1

GAIT-BASED HUMAN IDENTIFICATION FROM A MONOCULAR VIDEO SEQUENCE

Amit Kale

Center for Visualization and Virtual Environments
1, Quality St Suite 800-B
KY 40507 USA
E-mail: amit@cs.uky.edu

Aravind Sundaresan[†]

Amit K. RoyChowdhury

Department of Electrical Engineering
University of California, Riverside
CA 92521 USA
E-mail: amitrc@ee.ucr.edu

Rama Chellappa

† *Center for Automation Research*
University of Maryland at College Park
MD 20742 USA
E-mail:rama@cfar.umd.edu

Human gait is a spatio-temporal phenomenon that characterizes the motion characteristics of an individual. It is possible to detect and measure gait even in low-resolution video. In this chapter, we discuss algorithms for identifying people by their gait from a monocular video sequence. Human identification using gait, similar to text-based speaker identification, involves different individuals performing the same task and a template-matching approach is suitable for such problems. In situations where the amount of training data is limited, we demonstrate the utility of a simple width feature for gait recognition. By virtue of their deterministic nature, template matching methods have limited noise resilience. In order to deal with noise we introduce a systematic approach to gait recognition by building representations for the structural and dynamic components of gait using exemplars and hidden Markov models (HMMs). The above methods assume that an exact side-view of the subject is available in the probe sequence. For the case when the person walks at an arbitrary angle far away from the camera we present a view invariant gait recognition algorithm which is based on synthesizing a side view of a person from an arbitrary monocular view.

1. Introduction

Automated person identification is an important component of surveillance. An effective approach to person identification is to reduce it to the problem of identifying physical characteristics of the person. This method of identification of persons based on his/her physiological/behavioral characteristics is called biometrics. Established biometric methods range from fingerprint and hand-geometry techniques to more sophisticated methods based on face recognition and iris identification. Unfortunately, no single biometric is perfect or complete. Fingerprints and hand-geometry are reliable but require physical contact. Although, signatures based on face and iris are non-intrusive in nature, the applicability of all these methodologies is restricted to very controlled environments. In fact, current technology is capable of recognizing mostly frontal faces. At the time of writing, iris recognition is being attempted at distances of not more than five meters.

When person identification is attempted in natural settings such as those arising in surveillance applications, it takes on a new dimension. Biometrics such as fingerprint or iris are no longer applicable. Furthermore, night vision capability (an important component in surveillance) is usually not possible with these biometrics. Even though an IR camera would reveal the presence of people, the facial features are far from discernible in an IR image at large distances. A biometric that can address some of these shortcomings is human 'gait' or the walking style of an individual. The attractiveness of gait as a biometric arises from the fact that it is non-intrusive and can be detected and measured even in low resolution video. Furthermore, it is harder to disguise than static appearance features such as a face and it does not require a cooperating subject.

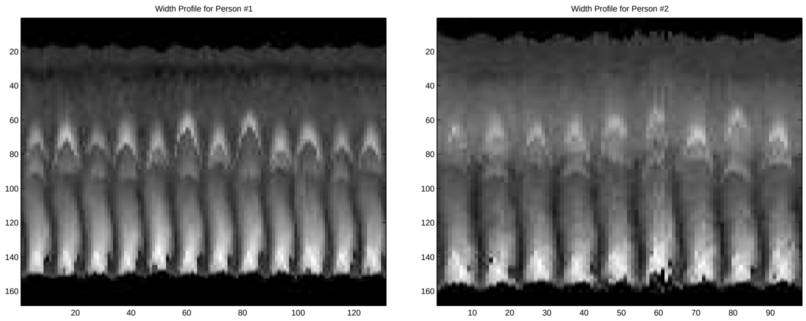
Early research on gait primarily involved psychophysical studies of gait viz. studying the ability of human observers to recognize gait. The belief that humans can distinguish between gait patterns of different individuals is widely held. Intuitively, it is possible to think of the qualities of walk such as stride length or body swing that help a perceiver identify an approaching figure even before the face becomes discernible. The earliest and most recognized psychophysical study of human perception of gait was the work of Johansson [1]. Small light bulbs were attached to the body joints of a darkly dressed walker. In this way only gait related cues were available and thus the perception of pure biological motion could be examined. When these point-light displays were static, the random collection of dots were variously interpreted as star constellations. However, as soon as the figures moved, the points of light were immediately perceived to be a human in motion. Motivated by Johanssons work, Kozlowski and Cutting [2] investigated whether observers could identify the gender of a point-light walker. The demonstration that gender could be extracted from gait provided insight into how observers might discriminate between gait patterns of different individuals. The prospect for observers being able to identify individuals from their gaits was thus encouraging. Cutting and Kozlowski [3] demonstrated that perceivers could reliably recognize themselves and their friends from dynamic point-light displays. Barclay et al. [4] suggested that individual walking styles might be captured by differences in a basic series of pendular limb motions.

Psychophysical evidence that there exists identity information in gait spurred development of computer vision based algorithms for gait-based human recognition. We attempt to give a summary of some of examples below, but the listing is by no means complete. Most of the methods for gait recognition are appearance based. Appearance based methods work reasonably well in the face of inaccurate background segmentation, changes in speed etc. However, such methods cannot tolerate drastic changes in clothing. Cunado et al. [5] extract a gait signature by fitting the movement of the thighs to an articulated pendulum-like motion model. Huang et al. [6] use optical flow to derive a motion image sequence for a walk cycle followed by eigenanalysis of the binarized silhouette to derive what are called eigen gaits. Benabdelkader et al. [7] use image self-similarity plots as a gait feature. Tolliver and Collins [8] use a spectral partitioning framework for identifying humans by their shape. Lee and Grimson [9] propose an approach in which several ellipses are fitted to different parts of the binarized silhouette of the person and the parameters of these ellipses such as location of its centroid, eccentricity etc. are used as a feature to represent the gait of a person. Hayfron-Acquah et al [10] proposed a method based on analyzing the symmetry of human motion using the Generalised Symmetry Operator. Han and Bhanu [11] proposed a gait-energy image approach for recognition.

2. Feature Selection

An important issue in gait is the extraction of appropriate salient features that will effectively capture the gait characteristics. The features must be reasonably robust to operating conditions and should yield good discriminability across individuals. As mentioned earlier, we assume that the side view of each individual is available. Intuitively, the silhouette appears to be a good feature to look at as it captures the motion of most of the body parts. It also supports night vision capability as it can be derived from IR imagery also. While extracting this feature we can either use the entire silhouette or use only the outer contour of the silhouette. The choice of using either of the above features depends upon the quality of the binarized silhouettes. If the silhouettes are of good quality, the outer contour retains all the information of the silhouette and allows a representation, the dimension of which is an order of magnitude lower than that of the binarized silhouette. However for low quality, low resolution data, the extraction of the outer contour from the binarized silhouette may not be reliable. In such situations, direct use of the binarized silhouette may be more appropriate.

We choose the width of the outer contour of the silhouette as one of our feature vectors. In order to generate the width vectors background subtraction [12] is first applied to the image sequence and the resulting motion image is binarized into foreground and background pixels. A bounding box is then placed around the part of the motion image that contains the moving person. Given the binarized silhouettes,



(a) (b)

Fig. 1. Width vector profile for several gait cycles of two individuals .

the left and right boundaries of the body are traced. The width along a given row is simply the difference in the locations of the right-most and the left-most boundary pixels in that row. It is easy to see that the norm of the width vector show a periodic variation. In Figure 1 we show plots of the width profiles of two different individuals for several gait cycles. Since we use only the distance between the left and right extremities of the silhouette, the two halves of the gait cycle are almost indistinguishable. From hereon, we refer to half cycles as cycles, for the sake of brevity. In Figure 1, the x-axis denotes the frame index while the y-axis denotes the index of the width vector (the row index). The ith horizontal line in the image shows the variations in the ith element of the width vector as a function of time. A brighter gray-scale indicates a higher value of the width. We observe that within each cycle, there is a systematic temporal progression of width vector magnitude, viz. the dynamics. A similar observation has been made in [13] where the gait patterns are analyzed as Frieze patterns. For the two width profile plots shown in Figure 1 , the differences are quite visible. For instance, by observing the bright patterns in the upper region of the two images we see that the brightness is more pronounced in the first image as compared to the second. This area of the plot corresponds to the swings of the hand. Secondly, note that the brightness gradient (which translates to velocity in the video sequence) in the lower part of the images is more pronounced for Person 1 as compared to Person 2. This part of the plot corresponds to the swings of the extremities of the foot. Additionally, note that the height, as well as the visibility of the neck part of the two persons are different. It must be pointed out, however, that the differences need not be so pronounced for all individuals. Thus, the width profile contains structural and dynamic information peculiar to each individual. Also, by definition, the width vector is translation-invariant. Hence, the width of the outer contour of the silhouette is indeed a potentially good candidate as a feature.

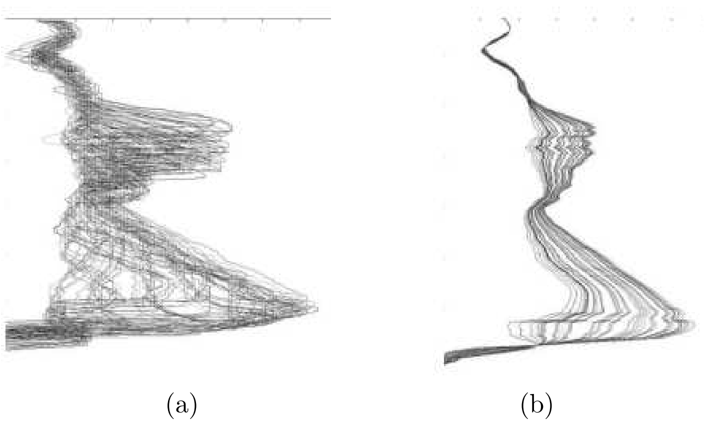
3. Gait-Based Human Identification Using Appearance Matching

A gait cycle corresponds to one complete cycle from rest (standing) position to-right-foot-forward-to-rest-to-left-foot-forward-to-rest position. The movements within a cycle consist of the motion of the different parts of the body such as head, hands, legs etc. The characteristics of an individual are reflected not only in the dynamics and periodicity of a gait cycle but also in the size and shape of that individual. Our aim is to build a model for representation and recognition of individual gait.

In a pattern classification problem, choice of the feature as well as the classifier is important. As discussed earlier, if the gait data is clean, the width of the outer contour of the silhouette of the person can be a good feature for gait recognition. From the temporal width plots, we note that although the width vector changes with time within a gait cycle, there is a high degree of correlation among the width vectors across frames. Most changes occur in the hand and in the leg regions. Hence, it is reasonable to expect that gait information in the width vector can be derived with much fewer coefficients. Given the width vectors $\{W(1), \cdots, W(N)\}$,for the N frames $W(.) \in R^M$, we compute the eigen vectors $\{V(1,) \cdots, V(M)\}$ corresponding to the eigen values of the scatter matrix arranged in the descending order and reconstruct the corresponding width vectors using $m(< M)$ most significant eigen vectors as

$$W_r(i) = (\sum_{j=1}^{m} w_j V(j)) + \bar{W}$$

where $w_j =< W(i), V(j) >$ and $\bar{W} = \frac{W(1)+\cdots+W(N)}{N}$. Figure 2 shows the width vectors reconstructed using two eigenvectors. Temporally-ordered sequences of eigensmoothed width vectors are used for compactly representing the person's gait.



(a) (b)

Fig. 2. Effect of eigen decomposition and reconstruction on the width vectors. (a) Overlapped raw width vectors (b) Smoothed width vectors. Notice that the leg region (the bottom half of the figures) contain a significant portion of the dynamics.

For a normal walk, gait sequences are repetitive and exhibit nearly periodic behavior. The gait problem is analogous to text-based speaker identification/verification wherein different individuals utter the same text but differ only in the characteristics of their utterance [14]. A template matching approach is suitable for such problems especially if the amount of training data is limited. Typically, gait cycles when taken at different times tend to be unequal in length due to changes in walking speeds of the individuals. To deal with this issue, dynamic time-warping (DTW) is employed for matching gait sequences. The DTW method was originally developed for isolated word recognition [15], and later adapted for text-dependent speaker verification [14]. DTW uses an optimum time expansion/compression function for producing non-linear time normalization so that it is able to deal with misalignments and unequal sequence lengths of the probe and the reference gait sequences. A distance metric (usually the Euclidean distance) defined as a function of time is computed between the two feature sets representing the gait data. A decision function is arrived at by integrating the metric over time. Assuming that the first frame of the reference and probe sequence are both indexed as 1 and the last frames of the reference and probe sequences be indexed as X and Y, respectively the match between the two sets can be represented by a sequence of K points $C(1), C(2), ..., C(k), ..., C(K)$, where $C(k) = (x(k), y(k))$, and $x(k)$ is a frame index of the probe sequence and $y(k)$ is a frame index of the reference sequence. Here, $C(k)$ represents a mapping of the time axis of probe sequence onto that of the reference sequence. The sequence $F = C(1), C(2), ...C(k), ..., C(K)$ is called the warping path. The process of time normalization involves computing the cumulative distance subject to endpoint, local continuity and global path constraints [16].

Table 1. Cumulative match scores for the UMD database using different eigen-features.

$Feature\backslash Rank$	1	2	3	4	5
Eigenvector 1	73	75	80	80	84
Eigenvectors 1,2	80	87	90	90	91
Eigenvectors 1,2,3	68	80	84	84	84
Eigenvectors 1,2,3,4	73	77	84	84	84

We experimented with different databases to test our method including the UMD, CMU and MIT datasets [17]. Table 1 shows the gait recognition result for the UMD dataset using different number of eigenvectors for reconstruction. Note that by using just the first two eigenvectors an accuracy of 80% is achievable. Other eigenvectors are noisy and, in fact, tend to lower the accuracy. We also considered the USF database[a] which has been identified as the gait challenge database [18]. The database has variations as regards viewing direction, shoe type, surface type.

[a] More details about this database can be found at http://figment.csee.usf.edu/GaitBaseline/

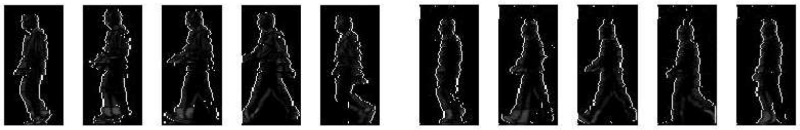
Also the subjects were asked to carry a briefcase for one testing condition. The USF database has the largest number of individuals among all the databases. It has variations with respect to floor surface (grass (G) or concrete(C)), shoe type (A or B), and camera viewing direction (left (L) or right (R)). The reference for all the experiments was chosen to be (G, A, R). The number of frames corresponding to four half cycles varied from 65 to 90. Different probe sequences for the experiments along with the cumulative match scores are given in Table 2 for the baseline algorithm [19] as well as our method using the eigensmoothed width feature. Note that recognition performance suffers most due to difference in surface characteristics, and least due to difference in viewing angle. An examination of the USF database revealed that the silhouettes provided were noisier compared to the previous datasets. We wanted to see what the performance would be by using the binary silhouettes directly as the feature. In this case we used the binary correlation distance in the local distance computation. As can be seen from the last two columns of Table 2 usage of the binarized silhouettes yields better performance numbers compared to the width vector in this case.

Table 2. Probe Sets and match scores for the USF database using the baseline algorithm and our approach using width feature and entire binary silhouette.

Experiment (Probe)	Baseline		Width Vector		Binary Silhouette	
	Rank 1	Rank 5	Rank 1	Rank 5	Rank 1	Rank 5
A (G, A, L)	79	96	79	91	84	97
B (G, B, R)	66	81	67	79	83	91
C (G, B, L)	56	76	30	55	59	79
D (C, A, R)	29	61	17	42	41	64
E (C, B, R)	24	55	15	39	24	53
F (C, A, L)	30	46	16	30	27	51
G (C, B, L)	10	33	9	31	24	38

4. A Framework for gait-based person identification using continuous HMMs

In the previous section we saw the application of a simple template matching approach to gait recognition. A careful analysis of gait would reveal that it has two important components. The first is a structural component that captures the physical build of a person e.g. body dimensions, length of limbs etc. The second component is the motion kinematics of the body during a gait cycle. A recent paper by Veeraraghavan et al [20] evaluates the contribution from these factors in vision-based gait recognition. In this section we propose a systematic approach to gait recognition by building representations for the structural and kinematic components of gait. A closer examination of the physical process behind the generation of gait signature reveals that, during a gait cycle, it is possible to identify certain distinct phases or stances. In Figure 3, we show five frames that we have picked from a gait cycle for two individuals.



(a) (b)

Fig. 3. Stances corresponding to the gait cycle of two individuals .

In the first stance, the person holds the two feet together. In the second, he is just about to start and his hand is slightly raised. In the third stance, the hands and the feet are separated, while in the fourth, the hands and feet are displaced to a maximum. Finally, in the fifth stance, the person is returning to the rest state. Clearly, every person transits among these successive stances as he/she walks. Although, these stances are generic, there exist differences not only in their image appearance based on the physical build of an individual but also in the way an individual transits across these stances as he/she walks which represents the gait kinematics of the individual. A reasonable way to build a structural representation for a person is to pick N exemplars (or stances) $\mathcal{E} = \{\mathbf{e}_1, \cdots, \mathbf{e}_N\}$ from the pool of images that will minimize the error in representation of all the images of that person. Given the image sequence for an unknown person $\mathcal{Y} = \{\mathbf{y}(1), \cdots, \mathbf{y}(T)\}$, these exemplars can be directly used for recognition as

$$ID = \arg\min_j \sum_{t=1}^{T} \min_{n\in\{1,\cdots,N\}} d(\mathbf{y}(t), \mathbf{e}_n^j),$$

where $\mathbf{y}(t)$ represents the image of an unknown person at the tth time instant, while $\mathbf{e}_n^j$ represents the nth exemplar of the jth individual. Note, however, that such a simple discrimination criterion is susceptible to failures not only due to noise but more importantly due to the presence of structural similarities among people in the database. To improve discriminability, the kinematics of the data must be exploited. A closer look at the gait cycle reveals that there is a temporal progression in the proximity of the observed silhouette to the different exemplars. Note that at the start of the gait cycle, a frame is closer to the first exemplar as compared to the other four. As time progresses, the frame will be closer to the second exemplar as compared to the others and so on. Underlying the proximity of the silhouette to the exemplars is a probabilistic dependence across the exemplars. This encompasses information about how long a gait cycle persists in a particular exemplar as well as the way in which the gait cycle transits from one exemplar to the other. For two people who are similar in physical build, this kinematic knowledge can be used to improve the recognition performance. Because the transitions are systematic, it is possible to model this probabilistic dependence by a Markov matrix as shown below.

$$A = [P(\mathbf{e}_i(t)|\mathbf{e}_j(t-1))] \tag{1}$$

for $i, j \in \{1, \cdots, N\}$. The matrix A encodes the kinematics in terms of state duration densities and transition probabilities. Often, in a practical situation, only a finite amount of training data is available and modeling can be difficult if the feature dimensionality is high. The dimension of the feature vectors described in the previous section is at least 100. Directly using the feature vectors to estimate the structure of the person and the kinematics of gait is clearly not advisable. We propose two different approaches to model the structure and kinematics of gait viz. an indirect approach and a direct approach. The choice of nomenclature will become apparent in the following discussion.

4.1. *Approach 1: Indirect Approach*

In this approach we pick N exemplars (or stances) $\mathcal{E} = \{\mathbf{e}_1, \cdots, \mathbf{e}_N\}$ from the pool of images that will minimize the error in representation of all the images of that person. different starting positions). In order to do this, we divide each gait cycle into N equal segments. We pool the image features corresponding to the ith segment for all the cycles. The centroids (essentially the mean) of the features of each part were computed and denoted as the exemplar for that part. Doing this for all the N segments gives the optimal exemplar set $\mathcal{E} = \{\mathbf{e}_1^*, \cdots, \mathbf{e}_N^*\}$. The next issue is how to choose N. In problems like image compression, it is a common practice to look at the rate-distortion curves to examine the marginal reduction in the distortion as the bits per pixel are increased. We found that increasing N beyond 5 or 6 does not lead to a significant drop in distortion.

In order to reliably estimate the gait kinematics we propose a novel way of compactly encoding the observations observations. Let $\mathbf{x}(t)$ denote the feature extracted from the image at time t. The distance of $\mathbf{x}(t)$ from the corresponding exemplars $\mathbf{e}_n \in \mathcal{E}$ can be computed to build a frame-to-exemplar distance (FED) vector, $\mathbf{f}(t)$, which serves as a lower N-dimensional representation of the image at time t. For instance, for the jth individual we compute the nth element of the FED vector as

$$[\mathbf{f}_j^{\mathcal{X}^j}(t)]_n = d(\mathbf{x}^j(t), \mathbf{e}_n^j), \tag{2}$$

where, $t \in \{1, \cdots, T\}$, $\mathbf{e}_n^j$ denotes the nth exemplar of the jth person and $n \in \{1, \cdots, N\}$. Thus, $\mathbf{f}_j^{\mathcal{X}^j}(t)$ constitutes an observation vector for person j. Similarly, $\mathbf{f}_j^{\mathcal{X}^i}(t)$ represents the observation sequence of the person i encoded in terms of the exemplars of person j. Note that for a frame at the start of the gait cycle, the first element of the observation vector will be smaller in magnitude as compared to the remaining four elements. As time progresses, the first element will increase in magnitude because the frame moves closer to the second stance. This temporal variation in the FED vector components corresponds precisely to the transition across exemplars. In particular it is possible to look upon the FED vector sequence $\mathbf{f}_j^{\mathcal{X}^j}(t)$ as the observed manifestation of the transition across exemplars (a hidden process). An HMM is appropriate for such a signal. HMMs [21] use a Markov process to model the changing statistical characteristics that are manifested in the actual observations. For the gait problem, the exemplars can be considered as analogues to the states of the HMM while the FED vector sequence can be considered as the observed process. Since the feature vectors are transformed to the intermediate FED vector representation, we refer to this approach as an indirect approach. In the proposed model for gait, the primary HMM parameters of interest are the number of states, the initial probability vector (π), the transition probability matrix (A) and the output probability distribution B which we model as a continuous probability distribution. $\lambda = (A, B, \pi)$ will be used to compactly represent the HMM.

In order to recognize an unknown person the FED vector sequence $\mathbf{f}_j^{\mathcal{Y}}(t)$ is computed for all $j \in \{1, \cdots, M\}$ for him/her using (2). The likelihood that the observation sequence $\mathbf{f}_j^{\mathcal{Y}}$ was generated by the HMM corresponding to the jth person can be deciphered by using the forward algorithm [21] as

$$P_j = \log(P(\mathbf{f}_j^{\mathcal{Y}} | \lambda_j)) \tag{3}$$

We repeat the above procedure for every person in the database thereby producing $P_j, j \in \{1, \cdots, M\}$. If the unknown person was m, P_m would be expected to be the largest among all P_j's since the distance between $\mathcal{Y}$ and the stances of person m will be smaller than that between $\mathcal{Y}$ and any other person. Also the pattern of transitions between stances/states for $\mathcal{Y}$ will be closest to that for person m.

4.2. *Approach 2: Direct Approach*

In this approach we use the feature vector in its entirety to estimate the HMM $\lambda = (A, B, \pi)$ for each person. Hence we refer to this approach as the direct approach. One of the important issues in training is learning the observation probability B. As discussed before, the reliability of the estimated B depends on the number of training samples available and the dimension of the feature vector. In order to deal with the high dimensionality of the feature vector, we propose an alternative representation for B. As discussed in the previous section it is possible, during a gait cycle, to identify certain distinct phases or stances. We build a structural representation for a person by picking N exemplars (or stances) from the training sequence, $\mathcal{X} = \{\mathbf{X}(1), \cdots, \mathbf{X}(T)\}$. We now define B in terms of the distance of this vector from the exemplars as follows.

$$b_n(\mathbf{X}(t)) = P(\mathbf{X}(t) | \mathbf{e}_n) = \beta e^{-\alpha D(\mathbf{X}(t), \mathbf{e}_n)} \tag{4}$$

The probability, $P(\mathbf{X}(t)|\mathbf{e}_n)$ is defined as a function of $D(\mathbf{X}(t), \mathbf{e}_n)$, the distance of the feature vector $\mathbf{X}(t)$ from the n^{th} exemplar, $\mathbf{e}_n$. The motivation behind using an exemplar-based model in the above manner is that the recognition can be based on the *distance measure* between the observed feature vector and the exemplars. During the training phase, a model is built for all the subjects in the gallery. Note that B is completely defined by $\mathcal{E}$ if α and β are fixed beforehand.

An initial estimate of $\mathcal{E}$ and λ is formed from $\mathcal{X}$, and these estimates are refined iteratively using Expectation-Maximization [22]. An initial estimate of an ordered set of exemplars from the sequence as described in Section 4.1. A corresponding initial estimate of the transition matrix, $A^{(0)}$ (with $A^{(0)}_{j,j} = A^{(0)}_{j,j \bmod N+1} = 0.5$, and all other $A^{(0)}_{j,k} = 0$) is also obtained. The initial probabilities $\pi^{(0)}_n$ are set to be equal to $1/N$. The iterative refinement of the estimates is performed in two steps. In the first step, a Viterbi evaluation [21] of the sequence is performed using the current values for the exemplars and the transition matrix. We can thus cluster feature vectors according to the most likely state they originated from. Using the current values of the exemplars, $\mathcal{E}^{(i)}$ and the transition matrix, $A^{(i)}$, Viterbi decoding on the sequence $\mathcal{X}$ yields the most probable path $\mathcal{Q} = \{q^{(i)}(1), q^{(i)}(2), \ldots, q^{(i)}(T)\}$, where $q^{(i)}(t)$ is the estimated state at time t and iteration i. Thus the set of observation indices, whose corresponding observation is estimated to have been generated from state n is given by $\mathcal{T}^{(i)}_n = \{t : q^{(i)}(t) = n\}$. The updated values of exemplars can be shown to be:

$$\mathbf{e}^{(i+1)}_n = \textstyle\sum_{t \in \mathcal{T}^{(i)}_n} \tilde{\mathbf{X}}(t) \qquad (5)$$

Given $\mathcal{E}^{(i+1)}$ and $A^{(i)}$, we can calculate $A^{(i+1)}$ using the Baum-Welch algorithm [21]. We set $\pi^{(i+1)}_n = \frac{1}{N}$ at each iteration. Thus we can iteratively refine our estimates of the HMM parameters. It usually takes only a few iterations to obtain an acceptable estimate.

Given the feature vector sequence of the unknown person, $\mathcal{Y}$, and the exemplars and HMM model parameters for the different people in the database, the likelihood that the observation sequence was produced by the jth individual in the database is computed using the forward algorithm as

$$P_j = \log(P(\mathcal{Y}|\lambda_j)). \qquad (6)$$

Note that λ_j implicitly includes the exemplar set corresponding to person j. The difference between the direct and indirect methods is that in the former the feat ure vector is directly used as the observation vector for the HMM whereas in the latter, the FED is used as the observation vector. We present the results of both our methods and a comparative analysis on the USF dataset. Different probe sequences for the experiments along with the cumulative match scores are given in Table 3 for the baseline algorithm [19], our direct and indirect approaches. The image quality for the USF database is worse than the previous two databases in terms of resolution and amount of noise. We experimented with both the width feature as well as the binarized silhouette for the USF dataset. However, the extraction of the outer contour in this case is not reliable and the width vectors were found to be noisy. In Table 3, we report only the results of our methods using the silhouettes as the image feature. From Table 3 we observe that the direct method is more robust to the presence of noise than the indirect method. We also note that the recognition performance suffers most due to differences in surface and background characteristics, and least due to difference in viewing angle. Results from other research groups using this data can be found in [8] and websites (http://degas.umiacs.umd.edu/links.html). From

Tables 2 and 3, we note that the HMM approach indeed surpasses the performance using the appearance matching method. Recently, the gallery in the USF database was extended by adding subjects who walked with only one shoe type on grass, which happened to be labelled as Shoe B. Since the shoe type labeling is arbitrary, they were put in the gallery to increase the gallery size to 122. The results for this case are reported in [23].

Table 3. Probe Sets and match scores for the USF database using the baseline algorithm and our indirect and direct approaches.

Experiment (Probe)	Baseline		Indirect Approach		Direct Approach	
	Rank 1	Rank 5	Rank 1	Rank 5	Rank 1	Rank 5
A (G, A, L)	79	96	91	100	99	100
B (G, B, R)	66	81	76	81	89	90
C (G, B, L)	56	76	65	76	78	90
D (C, A, R)	29	61	25	61	35	65
E (C, B, R)	24	55	29	39	29	65
F (C, A, L)	30	46	24	46	18	60
G (C, B, L)	10	33	15	33	24	50

5. View Invariant Gait Recognition

The gait of a person is best reflected when he/she presents a side view (referred to in this chapter as a canonical view) to the camera. Hence, most of the above gait recognition algorithms rely on the availability of the side views of the subject. The situation is analogous to face recognition where it is desirable to have frontal views of the person's face. In realistic scenarios, however, gait recognition algorithms need to work in a situation where the person walks at an arbitrary angle to the camera. For a person walking along a non-canonical direction, appearance based features which are used for recognition get distorted. To explain this better we consider the width feature discussed earlier. Temporal plots of the width-vector for the same person walking in the canonical and non-canonical($\theta = 45$) direction are shown in Figures 5 (a) and (b) respectively. A simple gait feature, viz. the stride length or the maximal separation of the feet, can be derived from the width plots by measuring the highest intensity in the leg regions(lower halves of the width plot). Clearly the apparent stride-length is smaller for the non-canonical view. The second effect that is obvious from the plots is a foreshortening effect as the person walks away from the camera. In order to obtain good gait recognition performance, it is necessary to correct for both of these effects through view synthesis. The most general solution to this problem is to estimate the 3-D model for the person. Features extracted from the 3-D model can then be used to provide the gait model for the person. This problem requires the solution of the structure from motion (SfM) or stereo reconstruction problems [24,25], which are known to be hard for articulating objects. In the absense of methods for recovering accurate 3-D models, a simple way to exploit existing appearance based methods is to synthesize the canonical views of a walking person. In [26], Shakhnarovich et al. compute an image based visual hull from a set of monocular views which is then used to render virtual canonical views for tracking and recognition. Gait recognition is achieved by matching a set of image features based on moments extracted from the silhouettes of the synthesized probe video to the gallery. An alternative to synthesizing canonical views is the work of Bobick and Johnson [27]. In this work, two sets of activity-specific static and stride parameters are extracted for different individuals. The expected confusion for each set is computed to guide the choice of parameters under different imaging conditions (viz. indoor vs outdoor, side-view vs angular-view etc). A cross-view mapping function is used to account for changes in viewing direction. The set of stride parameters (which is smaller than the set of static parameters) is found to exhibit greater resilience to viewing direction. Representation using such a small set of parameters may not give good recognition rates on large databases.

In this section we present a view-invariant gait recognition algorithm for the single camera case. Consider a person walking along a straight line which subtends an angle θ with the image plane (AC in Figure 4). If the distance, Z_0, of the person from the camera is much larger than the width, ΔZ, of the person, then it is reasonable to replace the scaling factor $\frac{f}{Z_0+\Delta Z}$ for perspective projection by an average scaling factor $\frac{f}{Z_0}$. In other words, for human identification at a distance, we can approximate the actual 3-D human as a planar object. Assume that we are given a video of a person walking at a fixed angle θ with a translational velocity $\mathbf{V} = [v_X, 0, v_Z]^T$(Figure 4). We show that by tracking the direction of motion, α, in the video sequence, we can estimate the 3-D angle θ. Using the planarity assumption, knowing angle θ and the calibration parameters, we can synthesize side-views of the sequence of images of an unknown walking person without explicitly computing the 3D model of the person.

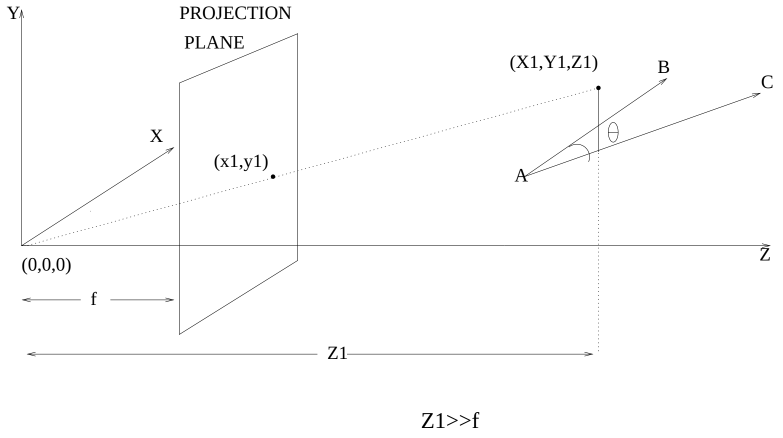


Fig. 4. Imaging Geometry

Tracking

Assuming that we can find the location (x_{ref}, y_{ref}) of the persons head at the start of such a segment, we use a sequential Monte Carlo particle filter [28] to track the head of the person to get $\{(x^i(t), y^i(t)), w^i(t)\}$ where the superscript denotes the index of the particle and $w^i(t)$ denotes the probability weight for the estimate $((x^i(t), y^i(t))$.

Estimation of 3-D Azimuth Angle

Assume that the motion between two consecutive frames in the video sequence is small. Using the optical flow based SfM equations [29] for constant velocity models $v_Z(t) = v_Z(\neq 0)$ and $v_X(t) = v_X(\neq 0)$, $cot(\theta(t)) = \frac{v_X}{v_Z}$ and given the initial position of the tracked point (x_{ref}, y_{ref}) it can be shown that

$$cot(\theta) = \frac{x_{ref} - y_{ref} cot(\alpha(x_{ref}, y_{ref}))}{f}, \tag{7}$$

Knowing (x_0, y_0),$cot(\alpha)$ and θ, f can be computed as part of a calibration procedure.

View Synthesis

Having obtained the angle θ, we synthesize the canonical view. Let Z denote the distance of the object from the image plane. If the dimensions of the object are small compared to Z, then the variation in θ, $d\theta \approx 0$. This essentially corresponds to assuming a planar approximation to the object. Let $[X_\theta, Y_\theta, Z_\theta]'$ denote the coordinates of any point on the person (as shown in the Figure 4) who is walking at an angle $\theta \geq 0$ to the plane passing through the starting point $[X_{ref} Y_{ref} Z_{ref}]'$ and parallel to the image plane which we shall refer to, hereafter, as the canonical plane. Computing the 3-D coordinates of the synthesized point involve a rotation about the line passing through the starting point followed by a perspective transformation we can obtain the equations for $[x_0, y_0]'$ as

$$\begin{aligned} x_0 &= f \frac{x_\theta cos(\theta) + x_{ref}(1 - cos(\theta))}{-sin(\theta)(x_\theta + x_{ref}) + f} \\ y_0 &= f \frac{y_\theta}{-sin(\theta)(x_\theta + x_{ref}) + f}, \end{aligned} \tag{8}$$

where $x = f\frac{X}{z}$ and $y = f\frac{Y}{z}$ (8) is attractive since it does not involve the 3D depth; rather it is a direct transformation of the 2D image plane coordinates in the non-canonical view to get the image plane coordinates in the canonical one. Thus using the estimated azimuth angle θ we can obtain a synthetic canonical view using (8). View synthesis provides for a correction of both the foreshortening and distortion of stride length (see Figures 5 (c) and (d)) and improves the gait recognition performance appreciably.

5.1. *Gait-based Recognition*

We present gait recognition results on two databases.

UMD3 database: This consists of 12 people, who walk along straight lines at different values of azimuth angle $\theta = 0, 15, 30$ and 45. The image sequences

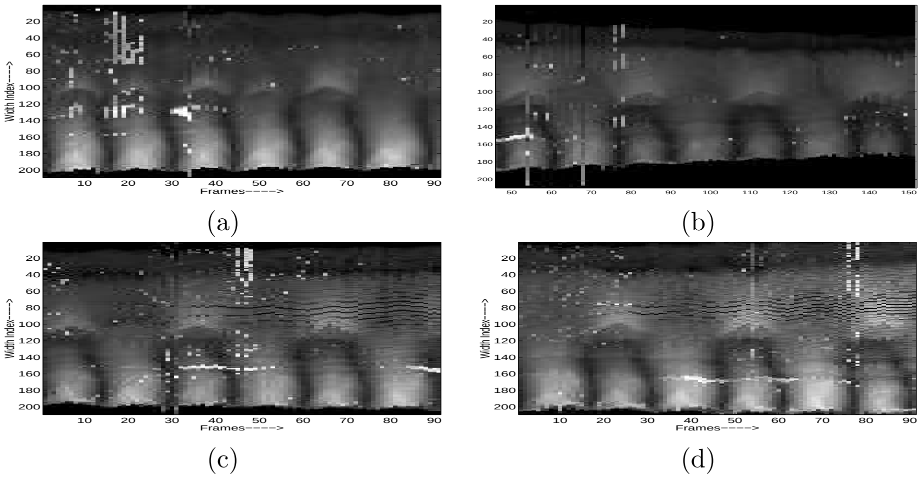
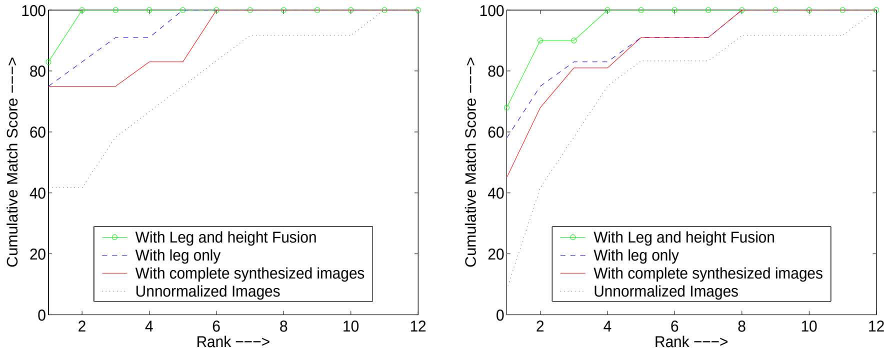


Fig. 5. Width profile as a function of time for (a) Canonical View ($\theta = 0$); (b)Unnormalized sequence for $\theta = 45$; synthesized views for: (c) $\theta = 30$ (d) $\theta = 45$.

corresponding to $\theta = 0$ were used as the gallery while the other sequences were used as a probe. The width profile plot for the canonical view and the view synthesized from $\theta = 45$ are shown in Figure 5. As can be seen from this plot, our method has compensated for both the foreshortening effect as well as restored the true leg-swing. Two consecutive cycles in the canonical view are chosen as the gallery to be compared with two consecutive gait cycles in the probe sequence. The DTW technique is used to match a given probe sequence to the different gallery sequences using binary correlation as a local distance measure and a similarity matrix $S = s(i,j)$ is obtained, where $s(i,j)$ refers to the similarity between the probe i and the gallery j. Gait recognition performance for $\theta = 30$ and 45^0 is shown in Figures 6(a) and (b) using the synthesized and raw images in terms of a cumulative match characteristic. As noted before, the algorithm results in a broader reproduction of the torso region. The situation can be remedied by assigning a lower weight to the torso region when computing the binary correlation or simply ignoring it. We take the latter approach by computing the binary correlation only over the lower half of the boxed image. The result using only the leg region is shown as the dashed lines in the Figure 6. It can be seen that the gait recognition result is better than what is obtained by using the entire body. Interestingly, [19] notes that the lower 20 % of the silhouette accounts for roughly 90% of the recognition. To boost the gait recognition performance further, certain structural characteristics of the individual that are extracted subsequent to view synthesis e.g. height can be fused with the leg dynamics. The height of the probe sequence is estimated robustly from the synthesized video as $h(i) = \text{median} h^j(i), j = 1 \cdots M$ M, being the length of the probe sequence. We fuse height information together with the leg dynamics by scaling each entry $s(i,j)$ of the similarity matrix by the corresponding height ratio, $\max(\frac{h(i)}{h(j)}, 2 - \frac{h(i)}{h(j)})$. The results for this case are shown as the solid line with circles in Figure 6. The fact that the gait recognition results are encouraging upto angles of 45 degrees allows us to hypothesize that it is possible to do reasonable human identification using gait with only two cameras (installed perpendicular to each other).



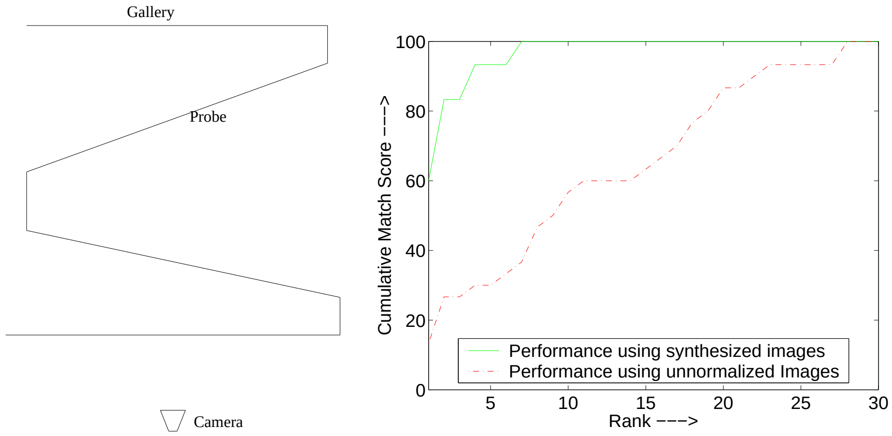
(a) (b)

Fig. 6. Cumulative Match Characteristics for Original and Synthesized images for (a) $\theta = 30$ (b) $\theta = 45$ for UMD3 database.

NIST database: This consists of 30 people walking along a Σ -shaped walking pattern as shown in Figure 7(a). There are two cameras looking at the top horizontal part of the sigma. The camera that is located further away is used in our experiments since the planar approximation we make is more valid in that case. The segment of the sigma next to the top horizontal part is used as a probe. This segment is at an angle 33^0 to the horizontal part. To do gait recognition we employed the fusion of the leg-dynamics with the height since it gave the best performance for Database 1. The gait recognition result is shown in Figure 7(b). As can be seen the recognition rate is about 60%. One of the reasons for the lower recognition performance in this case is that the image size is rather small. Note however that the recognition goes to 100% within 6 ranks.

6. Conclusions and Future Work

In this chapter we investigated the information contained in the video sequences of human gait and how to extract and represent that information in ways that facilitate human identification. Human identification using gait, similar to text-based speaker identification, involves different individuals performing the same task and a template-matching approach is suitable for such problems. In situations where the amount of training data is limited, we showed the utility of a simple feature viz. the width of the outer contour of the binarized silhouette of the subject and its



(a) (b)

Fig. 7. (a)Σ shaped walking pattern in the NIST database (b)Gait recognition performance on the NIST database.

derivatives for gait recognition in a dynamic time warping framework. By virtue of their deterministic nature, template matching methods have limited noise resilience. To improve robustness a systematic approach to gait recognition by building representations for the structural and dynamic components of gait using exemplars and hidden Markov models (HMMs) was discussed. Gait can serve as a cue for recognizing people if the database is small. But for large databases, gait information, by itself, may not be sufficient to recognize an individual. In fact, we must realize that the gait recognition capability of even humans is limited. However, it can be used to narrow down the list of potential matches. In a recent paper [30] we demonstrated the use of gait as a filter to achieve faster human identification by limiting the number of candidates being passed to a more accurate face recognition algorithm. Gait can also be used in conjunction with other cues such as the color of clothing etc. for short time verification problems viz "was this the same person who walked in front of this camera t minutes ago?". Finally, a view invariant gait recognition algorithm which is based on synthesizing a side view of a person from an arbitrary monocular view was discussed. We also presented a view invariant gait recognition algorithm. The fact that the gait recognition results are encouraging upto angles of 45 degrees allows us to hypothesize that it is possible to do reasonable human identification using gait with only two cameras (installed perpendicular to each other). This could prove to be less restrictive than the visual hull approach that needs at least 4 cameras. For indoor multiple camera settings it would also be interesting to study the contribution of 3-D information for gait recognition using multi-camera kinematic models [31,32].

References

1. G Johansson, "Visual motion perception," *Scientific American*, vol. 232, pp. 76–88, 1975.
2. L. Kozlowski and J. Cutting, "Recognizing the sex of a walker from a dynamic point display," *Perception and Psychophysics*, vol. 21, pp. 575–580, 1977.
3. J. Cutting and L. Kozlowski, "Recognizing friends by their walk:gait perception without familiarity cues," *Bulletin of the Psychonomic Society*, vol. 9, pp. 353–356, 1977.
4. C.D. Barclay, J. E. Cutting, and L.T. Kozlowski, "Temporal and spatial factors in gait perception that influence gender recognition," *Perception and Psychophysics*, vol. 23, pp. 145–152, 1978.
5. D. Cunado, J.M. Nash, M.S. Nixon, and J. N. Carter, "Gait extraction and description by evidence-gathering," *Proc. of the International Conference on Audio and Video Based Biometric Person Authentication*, pp. 43–48, 1995.
6. P.S. Huang, C.J. Harris, and M.S. Nixon, "Recognizing humans by gait via parametric canonical space," *Artificial Intelligence in Engineering*, vol. 13, no. 4, pp. 359–366, October 1999.
7. R. Cutler C. Benabdelkader and L.S. Davis, "Motion based recognition of people in eigengait space," *Proceedings of the IEEE Conference on Face and Gesture Recognition*, pp. 267–272, 2002.
8. D. Tolliver and R. Collins, "Gait shape estimation for identification," *Proceedings of AVBPA*, pp. 734–742, 2003.
9. L. Lee and W.E.L. Grimson, "Gait analysis for recognition and classification," *Proceedings of the IEEE Conference on Face and Gesture Recognition*, pp. 155–161, 2002.
10. J. Hayfron Acquah, M.S. Nixon, and J.N. Carter, "Automatic gait recognition by symmetry analysis," *Pattern Recognition Letters*, pp. 2175–2183, 2003.
11. J. Han and B. Bhanu, "Statistical feature fusion for gait-based human recognition," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.
12. A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background subtraction," *FRAME-RATE Workshop, IEEE*, 1999.
13. R.T. Collins Y. Liu and Y. Tsin, "Gait sequence analysis using frieze patterns," *Proceedings of the European Conference on Computer Vision*, vol. 2, pp. 657–671, 2002.
14. Sadaoki Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 2, pp. 254–272, April 1981.
15. H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-26 no. 1, pp. 43–49, 1978.
16. Jinu Mariam Zacharia, "Text-independent speaker verification using segmental, suprasegmental and source features," M.S. thesis, IIT Madras Chennai India, March 2002.
17. A. Kale, N. Cuntoor, B. Yegnanarayana, A.N. Rajagopalan, and R. Chellappa, "Gait-based human identification using appearance matching," in *Optical and Digital Techniques for Information Security*, B. Javidi, Ed. Springer Verlag, December 2004.
18. P. J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. W. Bowyer, "The gait identification challenge problem: Data sets and baseline algorithm," *Proc of the International Conference on Pattern Recognition*, 2002.
19. P. J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. W. Bowyer, "Baseline results for the challenge problem of human id using gait analysis," *Proc. of the 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2002.
20. A. Veeraraghavan, A. RoyChowdhury, and R. Chellappa, "Role of shape and kinemat-

ics in human movement analysis," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

21. L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, February 1989.
22. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.
23. A. Kale, A. Sundaresan, A. N Rajagopalan, N. Cuntoor, A. RoyChowdhury, V.Kruger, and R. Chellappa, "Identification of humans using gait," *IEEE Transactions on Image Processing*, July 2004.
24. O.D. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, MIT Press, 1993.
25. R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000.
26. G.Shakhnarovich, L.Lee, and T.Darrell, "Integrated face and gait recognition from multiple views," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, December 2001.
27. A.F. Bobick and A. Johnson, "Gait recognition using static activity-specific parameters," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001.
28. Michael Isard and Andrew Blake, "Contour tracking by stochastic propagation of conditional density," *Proceedings of ECCV*, , no. 1, pp. 343–356, 1996.
29. Vishwjit Nalwa, *A Guided Tour of Computer Vision*, Addison-Wesley, 1993.
30. A. Kale, A. RoyChowdhury, and R. Chellappa, "Fusion of gait and face for human identification," *Proc. of the ICASSP*, 2004.
31. A. Sundaresan, A. RoyChowdhury, and R. Chellappa, "3d modelling of human motion using kinematic chains and multiple cameras for tracking," *Proc. of Eighth International Symposium on the 3-D Analysis of Human Movement, Tampa*, 2004.
32. C. Bregler and J. Malik, "Tracking people with twists and exponential maps," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 1998.