

Human Identification Using Gait and Face

Rama Chellappa
Center for Automation Research
University of Maryland
College Park, MD 20740

Amit K. Roy-Chowdhury
Dept. of Electrical Engineering
University of California
Riverside, CA 92521

Shaohua Kevin Zhou
Center for Automation Research
University of Maryland
College Park, MD 20740

1 Introduction

One of the main goals of computer vision research is to develop methods for recognition of objects and events. A subclass of these problems is the recognition of humans and their activities. In this chapter, we summarize some of the recent methods developed for human recognition using face and gait.

Gait recognition is related to the broader problem of human motion modeling, which has very important implications for different areas like surveillance, medical diagnosis, entertainment industry, video communications, etc. Traditionally, there has been a keen interest in studying human motion in various disciplines. In psychology, Johansson conducted classic experiments by attaching light displays to various body parts and showed that humans can identify motion when presented with only a small set of these moving dots [32]. Muybridge captured the first photographic recordings of humans and animals in motion in his famous publication on animal locomotion towards the end of the 19th century [52]. In kinesiology the goal has been to understand human motion with applications in sports, medicine, elderly care and early detecting of movement disorders [28]. Gait recognition is a relatively new area to computer vision researchers. However, significant progress has been made and reasonably good performance on large datasets under controlled circumstances has been achieved. The problems facing this area are poor performance in uncontrolled outdoor situations and the effects of time. Some progress has also been made in recognizing people walking in arbitrary directions to the camera.

The problem in face recognition can be defined as follows. A database of a large number of faces is available as the gallery. The faces may be represented as a single image or a set of images, either as a video sequence or a collection of discrete poses. These images are usually referred to as training images, since they are used to train the parameters of a recognition algorithm. Given an image or a set of images of an individual (known as test images), the problem is to identify the individual from the gallery or decide that he/she is not part of the gallery. The main challenges in face recognition are:

- Varying conditions of illumination between training and test images.
- Changes in appearance, make-up and clothing between training and test images.
- Changes due to difference in time between the recording of training and test images.
- Different poses of the face in different instances of recording.

All of these issues make face recognition an extremely challenging problem. However, considerable progress has been made in the last decade and face recognition technologies are under consideration for deployment at various public facilities.

2 Review of Existing Work

2.1 Human Recognition Using Gait

Human gait is a spatio-temporal phenomenon that characterizes the motion of an individual. When person identification is attempted in natural settings such as those arising in surveillance applications, 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. 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 face and it does not require a cooperating subject.

Although study of human gait is a relatively new area for computer vision researchers, extensive work has been carried out in the psychophysics community on the ability of humans to recognize others by their style of walking. In the computer vision community, research on gait has concentrated mostly on recognition algorithms. These methods can be divided into two groups - appearance based and model based. Appearance based methods can be further divided as deterministic or stochastic methods. Deterministic appearance based methods are [55, 35, 10], while well-known stochastic methods in the same category use Hidden Markov Models (HMM) [70, 41]. Model based methods are fewer largely because of the difficulty of obtaining accurate 3D models of the human body.

The belief that humans can distinguish between gait patterns of different individuals is widely held. These gait related quantities include stride length, bounce, rhythm, speed and perhaps even attributes such as swagger or body swing. The suggestion that humans can identify people by their gait was investigated in a series of early studies by Johansson [33]. He presented participants with images that had been reduced to point-light displays. His experiments suggested that we have some implicit notion of human movement, and can recognize temporal data within this context. Later work using point-light displays went further, demonstrating that not only could a walking figure rapidly be extracted from the moving lights, but also a perceiver could distinguish between different sorts of biological motion including walking, climbing stairs, jumping etc [15]. Attempts to address the question of identification from gait have proceeded in small steps. Kozlowski and Cutting [38] first investigated whether perceivers could identify the gender of a point-light walker. Their results indicated an accuracy rate of 65% and 70% when the walker was viewed from the side. In [14], it was suggested that gender may be identified indirectly through a determination of the "center of moment" of a walker. The demonstration that gender could be extracted from gait provided insight into how perceivers might discriminate between gait patterns of different individuals. Cutting and Kozlowski [13] demonstrated that perceivers could reliably recognize themselves and their friends from dynamic point-light displays. Barclay et al. [2] suggested that individual walking styles might be captured by differences in a basic series of pendular limb motions. Interestingly, Beardsworth and Buckner [4] have shown that the ability to recognize oneself from a point-light display is greater than the ability to recognize one's friends, despite the fact that we rarely see our own gait from a third-person perspective. Stevenage et al. [69] also explored the ability of people to identify others using gait information alone. He found that even with a brief exposure time and unfamiliarity with the walking subjects, the perceivers could identify the target correctly at greater than chance rate.

A recent study by Schollhorn et al. [62] studied the gait of fifteen subjects to study the presence of identity information in gait. It was found in this study that kinetic variables (captured using a force platform) as well as kinematic variables (captured by reflective markers on the thigh, shank and hip) were both necessary for gait identification. Furthermore simply using the leg portion of the body was adequate for getting good identification

performance.

Approaches to gait recognition problem can be broadly classified as being either model-based or model-free. Both methodologies follow the general framework of feature extraction, feature correspondence and high-level processing. The major difference is with regard to feature correspondence between two consecutive frames. Methods which assume *a priori* models match the 2-D image sequences to the model data. Feature correspondence is automatically achieved once matching between the images and the model data is established. Examples of this approach include the work of Lee et al. [42], where 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. Recognition is achieved by template matching. In [12], Cunado et al. extract a gait signature by fitting the movement of the thighs to an articulated pendulum-like motion model. The idea is somewhat similar to an early work by Murray [51] who modeled the hip rotation angle as a simple pendulum, the motion of which was approximately described by simple harmonic motion. Model-free methods establish correspondence between successive frames based upon the prediction or estimation of features related to position, velocity, shape, texture and color. Alternatively, they assume some implicit notion of what is being observed. Examples of this approach include the work of Huang et al. [31], where optical flow is used to derive a motion image sequence for a walk cycle. Principal components analysis is then applied to the binarized silhouette to derive what are called eigen gaits. Benabdelkader et al. [8] use image self-similarity plots as a gait feature. Little and Boyd [45] extract frequency and phase features from moments of the motion image derived from optical flow and use template matching to recognize different people by their gait. A dynamic time warping (DTW) [22] based algorithm for gait recognition was proposed in [35]. The algorithm matches two gait sequences (probe and gallery) by computing the distance, as a function of time, between two feature sets representing the data. This approach can be used even when substantial training data is not available. It can also account for modest variation in speed of walking. Two of the most successful approaches, till date, in gait recognition are [70] and [72]. In [70], the authors used a HMM [58] to represent the gait of each individual. This algorithm will be described in detail in the next section. A method for identifying individuals by shape, which is automatically extracted from a cluster of similar poses obtained from a spectral partitioning framework, was proposed in [72]. Most of the above methods rely on the availability of a side view in order to extract the gait parameters. Two approaches that address the problem of view-invariant recognition are [27] and [34]. A study of the role of kinematics and shape in computer vision based gait recognition problems was presented in [75].

Similar to the FERET evaluations for face recognition, a HumanID Gait Challenge Problem was introduced in order to measure progress of different recognition algorithms [56] (<http://www.gaitchallenge.org>). The challenge problem consists of a baseline algorithm, a set of twelve experiments and a dataset of 122 people. The baseline algorithm estimates silhouettes by background subtraction, and performs recognition by temporal correlation of silhouettes. Twelve experiments examine the effects of five covariates: change of viewing angle, change in shoe type, change in walking surface, carrying or not carrying a briefcase, and temporal differences. A description of the different experiments (probe sets) is given in Table 1. Identification and verification scores for all the experiments are reported using the baseline algorithm. The relative performance of the HMM based method with the baseline will be presented in the next section.

2.2 Human Recognition Using Face

Chronologically speaking, face recognition first started with still images. Popular methods that have been proposed are principal components analysis or eigenfaces [73, 50], linear discriminant analysis or Fisherfaces [80, 5], elastic graph matching [77], local feature analysis [53], morphable models [6], and numerous others. The reader is referred

Description of GaitChallenge Data

Experiment	Probe Description (Surface C/G, Shoe A/B, Camera L/R, Carry NB/BF, Time)	Number of Subjects
A	(G,A,L,NB,T1+T2)	122
B	(G,B,R,NB,T1+T2)	54
C	(G,B,L,NB,T1+T2)	54
D	(C,A,R,NB,T1+T2)	121
E	(C,B,R,NB,T1+T2)	60
F	(C,A,L,NB,T1+T2)	121
G	(C,B,L,NB,T1+T2)	60
H	(G,A,R,BF,T1+T2)	120
I	(G,B,R,BF,T1+T2)	60
J	(G,A,L,BF,T1+T2)	120
K	(G,A/B,R,NB,T2)	33
L	(G,A/B,R,NB,T2)	33

Table 1: Probe Sets for the GaitChallenge Data. The gallery is (G,A,R,NB,T1+T2). C/G represents concrete/grass surface, L/R represents left or right camera, NB/BF represents carrying a briefcase or not, T1 and T2 represent the data collected at two different time instants.

to a recent survey paper [81] for additional details on recent work on face recognition.

Statistical approaches to face modeling have been very popular since Turk and Pentland’s work on eigenface in 1991 [73]. In statistical approach, the two-dimensional appearance of face image is treated as a vector by scanning the image in lexicographical order, with the vector dimension being the number of pixels in the image. In the eigenface approach [73], all face images consists of a distinctive face subspace. This subspace is linear and spanned by the eigenvectors of the covariance matrix found using PCA. Typically we keep the number of eigenvectors much less than the true dimension of the vector space. The task of face recognition is then to find the closest match in this face subspace. However, PCA might not be efficient in terms of recognition accuracy since the construction of the face subspace does not capture class separability between humans. This motivated the use of LDA [5] and its variants. In LDA, the linear subspace is constructed in such a manner that the within-class scatter is minimized and the between-class scatter is maximized. This idea is further generalized in the approach called Bayesian face recognition [49], where intra-personal space (IPS) and extra-personal space (EPS) are used in lieu of within-class scatter and between-class scatter measures. The IPS models the variations in the appearance of the same individual and the EPS models the variations in the appearance due to a difference in the identity. Probabilistic subspace density is then fitted on each space. A Bayesian decision is taken using a *maximum a posteriori* (MAP) rule to determine the identity. In the famous EGM [77] algorithm, the face is represented as a labeled graph. The nodes of the graph are located at facial landmarks, e.g., the pupils, the tip of nose, etc. Also, each node is labeled with jets derived from responses obtained by convolving the image with a family of Gabor functions. Edges in the graph represent the geometric distance between two nodes. Face recognition is then formalized as a graph matching problem. All the above approaches are based on 2-D appearances and perform poorly when significant pose and illumination variations are present [81].

While recognition rates under controlled indoor situations are reasonably good, a lot needs to be done before such technologies can be deployed in outdoor situations. Many researchers believe that the use of video sequences,

as opposed to a single image, will lead to much better recognition rates. This is based on the intuition that integrating the recognition performance over a sequence would give a better result than considering just one single image from that sequence. Therefore, most of the present research in this area is in exploiting video sequence.

However, nearly all video-based recognition systems apply still-image-based recognition to selected good frames. In [30, 48, 76], RBF (Radial Basis Function) networks are used for tracking and recognition purposes. In [30], the system uses an RBF (Radial Basis Function) network for recognition. Since no warping is done, the RBF network has to learn the individual variations as well as possible transformations. The performance appears to vary widely, depending on the size of the training data. [76] presents a fully automatic person authentication system. The system uses video break, face detection, and authentication modules and cycles over successive video images until a high recognition confidence is reached. This system was tested on three image sequences; the first was taken indoors with one subject present, the second was taken outdoors with two subjects, and the third was taken outdoors with one subject in stormy conditions. Perfect results were reported on all three sequences, when verified against a database of 20 still face images. A multimodal based person recognition system is described in [9]. This system consists of a face recognition module, a speaker identification module, and a classifier fusion module. The most reliable video frames and audio clips are selected for recognition. 3D information about the head is used to detect the presence of an actual person as opposed to an image of that person. Recognition and verification rates of 100% were achieved for 26 registered clients.

3 Gait Recognition Using Hidden Markov Models

The HMM approach is suitable because the gait of an individual can be visualized as his/her adopting postures from a set, in a sequence which has an underlying structured probabilistic nature. The postures that the individual adopts can be regarded as the states of the HMM and are typical to that individual and provide a means of discrimination. This approach assumes that, during a walk cycle, the individual transitions among N discrete postures or states. An adaptive filter is used to automatically detect the cycle boundaries. The method is not dependent on the particular feature vector used to represent the gait information contained in the postures. The statistical nature of the HMM lends robustness to the model. In the method described below, the binarized background-subtracted image is used as the feature vector and different distance metrics, such as those based on the L_1 and L_2 norms of the vector difference, and the normalized inner product of the vectors, are used to measure the similarity between feature vectors.

3.1 Overview of the HMM Method

Let the database consists of video sequences of P persons. The model for the p^{th} person is given by $\lambda_p = (A_p, B_p, \pi_p)$ with N number of states. The model, λ_p , is built from the observation sequence for the p^{th} person using the sequence of feature vectors given by $\mathcal{O}_p = \{\mathbf{O}_1^p, \mathbf{O}_2^p, \dots, \mathbf{O}_{T_p}^p\}$, where T_p is the number of frames in the sequence of the p^{th} person. A_p is the transition matrix, and π_p is the initial distribution. The B_p parameter consists of the probability distributions for a feature vector conditioned on the state index, i.e., the set $\{P_1^p(\cdot), P_2^p(\cdot), \dots, P_N^p(\cdot)\}$. The probability distributions are defined in terms of *exemplars*, where the j^{th} exemplar is a typical realization of the j^{th} state. The exemplars for the p^{th} person are given by $\mathcal{E}_p = \{\mathbf{E}_1^p, \mathbf{E}_2^p, \dots, \mathbf{E}_N^p\}$. Henceforth, the superscript denoting the index of the person is dropped for simplicity. The motivation behind using an exemplar-based model is that the recognition can be based on the distance measure between the observed feature vector and the exemplars. The distance metric is evidently a key factor in the performance of the algorithm.

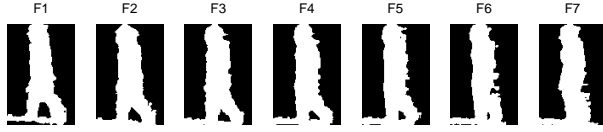


Figure 1: Part of an Observation Sequence

$P_j(\mathbf{O}_t)$ is defined as a function of $D(\mathbf{O}_t, \mathbf{E}_j)$, the distance of the feature vector \mathbf{O}_t from the j^{th} exemplar.

$$P_j(\mathbf{O}_t) = \alpha e^{-\alpha D(\mathbf{O}_t, \mathbf{E}_j)} \quad (1)$$

During the **training** phase, a model is built for all the subjects, indexed by $p = 1, 2, \dots, P$, in the gallery. An initial estimate of \mathcal{E}_p and λ_p is formed from \mathcal{O}_p , and these estimates are refined iteratively. Note that B is completely defined by \mathcal{E} if α is fixed beforehand. We can iteratively estimate A and π by using the Baum-Welch algorithm, keeping \mathcal{E} fixed. The algorithm to re-estimate \mathcal{E} is determined by the choice of the distance metric. During **testing**, given a Gallery $\mathcal{L} = \{\lambda_1, \lambda_2, \dots, \lambda_P\}$ and the probe sequence of length T , $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T\}$ traversing the path $\mathcal{Q} = \{q_1, q_2, \dots, q_T\}$, q_t being the state index at time t , we obtain the ID of the probe sequence as

$$ID = \arg_p \max_{\mathcal{Q}, p} \Pr[\mathcal{Q} | \mathcal{X}, \lambda_p]. \quad (2)$$

The feature vector used is the binarized version of the background subtracted images. The images are scaled and aligned to the center of the frame as in Figure 1 which features part of a sequence of feature vectors. We now describe the methods used to obtain initial estimates of the HMM parameters, the training algorithm and finally, identification results using USF data described in [56].

3.2 Initial Estimate of HMM Parameters

In order to obtain a good estimate of the exemplars and the transition matrix, we first obtain an initial estimate of an ordered set of exemplars from the sequence and the transition matrix and successively refine the estimate. The initial estimate for the exemplars, $\mathcal{E}^0 = \{\mathbf{E}_1^0, \mathbf{E}_2^0, \dots, \mathbf{E}_N^0\}$ is such that the only transitions allowed are from the j^{th} state to either the j^{th} or the $(j \bmod N + 1)^{\text{th}}$ state. A corresponding initial estimate of the transition matrix, A^0 (with $A_{j,j}^0 = A_{j, j \bmod N + 1}^0 = 0.5$, and all other $A_{j,k}^0 = 0$) is also obtained. The initial probabilities π_j are set to be equal to $1/N$.

We observe that the gait sequence is quasi-periodic and we use this fact to obtain the initial estimate \mathcal{E}^0 . We can divide the sequence into “cycles”, where a cycle is defined as that segment of the sequence bounded by silhouettes where the subject has arms by his/her side and legs approximately aligned with each other. We can further divide each cycle into N temporally adjacent clusters of approximately equal size. We visualize the frames of the j^{th} cluster of all cycles to be generated from the j^{th} state. Thus we can get a good initial estimate of \mathbf{E}_j from the feature vectors belonging to the j^{th} cluster. For example, assume that the training sequence is given by $\mathcal{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_T\}$. We can partition the sequence into K cycles, with the k^{th} cycle given by frames in the set $\mathcal{Y}_k = \{\mathbf{Y}_{S_k}, \mathbf{Y}_{S_k+1}, \dots, \mathbf{Y}_{S_k+L_k-1}\}$, where S_k and L_k are the index of the first frame of the k^{th} cycle, and the length of the k^{th} cycle respectively. We define the first cluster to comprise of frames with indices $S_k, S_k + 1, \dots, S_k + \frac{1}{2}L_k/N, S_k + L_k - \frac{1}{2}L_k/N, S_k + L_k - \frac{1}{2}L_k/N + 1, \dots, S_k + L_k - 1$. The j^{th} cluster ($j = 2, 3, \dots, N$) consists of frames with indices $S_k + (j - \frac{3}{2})L_k/N, S_k + (j - \frac{3}{2})L_k/N + 1, \dots, S_k + (j - \frac{1}{2})L_k/N$. We need to robustly estimate the cycle boundaries so that we can partition the sequence into N clusters and obtain the initial estimates of the exemplars. If the sums of the foreground pixels of each image are plotted with respect

to time, then, as per our definition of a cycle, the minima should correspond to the cycle boundaries. We denote the sum of the foreground pixels of the silhouette in the n^{th} frame as $s[n]$. This signal is noisy and may contain several spurious minima. However we can exploit the quasi-periodicity of the signal and filter the signal to remove the noise before identifying the minima. Methods such as median filtering or differential smoothing of $s[n]$ are not very robust as they do not take into account the frequency of the gait.

The specifications of the band-pass filter are such as to allow frequencies that are typical for a fast walk. The video is captured at 30 frames per second, and the sampling frequency, $f_s = 1/30$ and $T_s = 30$. The maximum gait frequency is assumed to be $f_m = 0.1$ corresponding to a cycle period of $T_m = 10$. A Hamming window of length L is used. The extended sequence $x[n]$ is obtained by symmetrically extending $s[n]$ in both directions by $L/2$. Therefore the sequence $x[n]$ has length $M = N + L$. The resultant sequence is filtered using a bandpass filter (with upper cut-off frequency $f_{uc} = f_m$), in both directions to remove phase delay. The distances between the minima of the filtered sequence provide an estimate of the cycle period. The cycle frequency is estimated as the inverse of the median of cycle periods. Using this revised estimate of the frequency of the gait, \hat{f} , a new filter is constructed with upper cut off frequency $f_{uc} = \hat{f} + 0.02$. A manual examination of all the sequences in the Gallery in the GaitChallenge database revealed a 100% detection rate with hardly any false detection of cycle boundaries.

3.3 Training the HMM Parameters

The iterative refining of the estimates is performed in two steps. In the first step, a Viterbi evaluation [58] of the sequence is performed using the current values for the exemplars and the transition matrix. Thus feature vectors are clustered according to the most likely state they originated from. The exemplars for the states are newly estimated from these clusters. Using the current values of the exemplars, $\mathcal{E}^{(i)}$ and the transition matrix, $A^{(i)}$, Viterbi decoding is performed on the sequence \mathcal{Y} to obtain the most probable path $\mathcal{Q} = \{q_1^{(i)}, q_2^{(i)}, \dots, q_T^{(i)}\}$, where $q_t^{(i)}$ is the state at time t . Thus the set of observation indices, whose corresponding observation is estimated to have been generated from state j is given by $\mathcal{T}_j^{(i)} = \{t : q_t^{(i)} = j\}$. We now have a set of frames for each state and we would like to select the exemplars so as to maximise the probability in (3). If we use the definition in (1), (4) follows.

$$\mathbf{E}_j^{(i+1)} = \arg_{\mathbf{E}} \max \prod_{t \in \mathcal{T}_j^{(i)}} P(\mathbf{Y}_t | \mathbf{E}) \quad (3)$$

$$\mathbf{E}_j^{(i+1)} = \arg_{\mathbf{E}} \min \sum_{t \in \mathcal{T}_j^{(i)}} D(\mathbf{Y}_t, \mathbf{E}) \quad (4)$$

The actual method for minimising the distance in (4) however depends on the distance metric used. We have experimented with three different distance measures, namely the Euclidean (EUCLID) distance, the inner product (IP) distance, and the sum of absolute difference (SAD) distance which are given by (5), (6), and (7) respectively. Note that though \mathbf{Y}_t and \mathbf{E} are 2-dimensional images, they are represented as vectors of dimension $D \times 1$ for ease of notation. $\mathbf{1}_{D \times 1}$ is a vector of D ones.

$$D_{EUCLID}(\mathbf{Y}, \mathbf{E}) = (\mathbf{Y} - \mathbf{E})^T (\mathbf{Y} - \mathbf{E}) \quad (5)$$

$$D_{IP}(\mathbf{Y}, \mathbf{E}) = 1 - \frac{\mathbf{Y}^T \mathbf{E}}{\sqrt{\mathbf{Y}^T \mathbf{Y} \mathbf{E}^T \mathbf{E}}} \quad (6)$$

$$D_{SAD}(\mathbf{Y}, \mathbf{E}) = |\mathbf{Y} - \mathbf{E}|^T \mathbf{1}_{D \times 1} \quad (7)$$

The equations for updating the j^{th} element of the exemplars in the EUCLID distance, IP distance and the SAD distance cases are presented in (8), (9) and (10) respectively. $\tilde{\mathbf{Y}}$ denotes the normalized vector \mathbf{Y} and $|\mathcal{T}_j^{(i)}|$

denotes the cardinality of the set $\mathcal{T}_j^{(i)}$.

$$\mathbf{E}_j^{(i+1)}(j) = \frac{1}{|\mathcal{T}_j^{(i)}|} \sum_{t \in \mathcal{T}_j^{(i)}} \mathbf{Y}_t(j) \quad (8)$$

$$\mathbf{E}_j^{(i+1)}(j) = \sum_{t \in \mathcal{T}_j^{(i)}} \tilde{\mathbf{Y}}_t(j) \quad (9)$$

$$\mathbf{E}_j^{(i+1)}(j) = \text{median}_{t \in \mathcal{T}_j^{(i)}} \{\mathbf{Y}_t(j)\} \quad (10)$$

Given $\mathcal{E}^{(i+1)}$ and $A^{(i)}$, we can calculate $A^{(i+1)}$ using the Baum-Welch algorithm [58]. Thus we can successively refine our estimates of the HMM parameters. It usually takes only a few iterations in order to obtain an acceptable estimate.

3.4 Identifying from a Test Sequence

Identifying a sequence involves deciding which of the model parameters to use for discrimination parameters. Given the models in the gallery, $\mathcal{L} = \{\lambda_1, \lambda_2, \dots, \lambda_P\}$ and the probe sequence, $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T\}$, we find the model and the path that maximizes the probability of the path given the probe sequence. The ID is obtained as in (2).

We do not need to use the trained parameter set, λ , as a whole. For example, if we believe that the transition matrix is predominantly indicative of the speed at which the subject walks, and is therefore not suitable as a discriminant of the ID of the subject, then we have the option of using only part of the parameter set given by $\gamma_p = (B_p, \pi_p)$ instead of using the HMM parameter set in its entirety. In this case, the conditional probability of the sequence, given the ID, is given as follows. The Baum-Welch Algorithm could be used in order to obtain $A_p^{\mathcal{X}}$ recursively in (12).

$$\Pr[\mathcal{Q}|\mathcal{X}, \gamma_p] = \Pr[\mathcal{Q}|\mathcal{X}, A_p^{\mathcal{X}}, \gamma_p] \quad (11)$$

$$A_p^{\mathcal{X}} = \arg_A \max \Pr[\mathcal{X}|A, \gamma_p] \quad (12)$$

3.5 Experimental Results

The objective of our experiments was to evaluate the performance of the algorithm and also compare the efficacy of the different distance measures in gauging the similarity between two images as far as posture is concerned. As described before, the GaitChallenge or USF database contains video sequences of 122 individuals, a subset of whom feature in sequences collected under each of 12 different conditions. The sequences are labeled Gallery and Probe A-L. We trained our parameters on the sequences from the Gallery set. In each experiment, we tried to identify the sequences in each of the seven Probe sets from the parameters obtained from the Gallery set using the inner product distance measure. The ID was calculated using (2). The experiments were repeated with different distance measures. The results of the experiment using the IP distance measure between feature vectors in the form of Cumulative Match Scores (CMS) plots [55] are in Figure 2(a). We observe that the distance measure that works best and is most simple to implement is the inner product distance. The performance comparison with the baseline [55] is illustrated in Figure 2(b).

From the experiments we note that the biggest drop in performance occurs due to change in surface type and when there is a difference in time between the gallery and the probe. Reasons for the sudden drop are not yet fully understood. Probable causes may be change in the silhouette (especially lower part for surface change) and change in clothing due to time differences. Note that the performance does not change much with small changes in viewing direction. In summary, gait recognition under arbitrary conditions is still an open research problem.

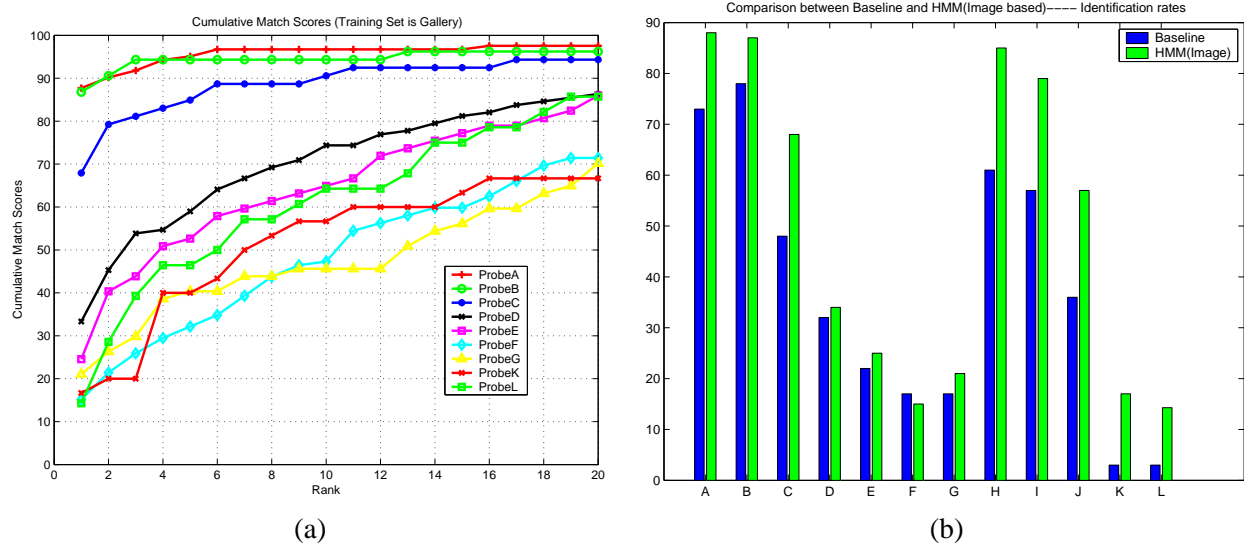


Figure 2: (a):CMS plots of Probes A-L tested against Gallery. (b): Comparison of identification rates of HMM and Baseline Algorithm.

4 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 view 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. 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 [18, 29], which are known to be hard for articulating objects. In the absence of methods for robust recovery of accurate 3-D models, a simple way to exploit existing appearance based methods is to synthesize the canonical views of a walking person. In [27], 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 [7]. 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. A method for recognizing the gait of an individual using joint angle trajectories was presented in [71]. However, representation using such a small set of parameters may not give good recognition rates on large databases.

We have developed a view-invariant gait recognition algorithm for the single camera case by synthesizing a canonical view from an arbitrary one without explicitly computing the 3-D depth. 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

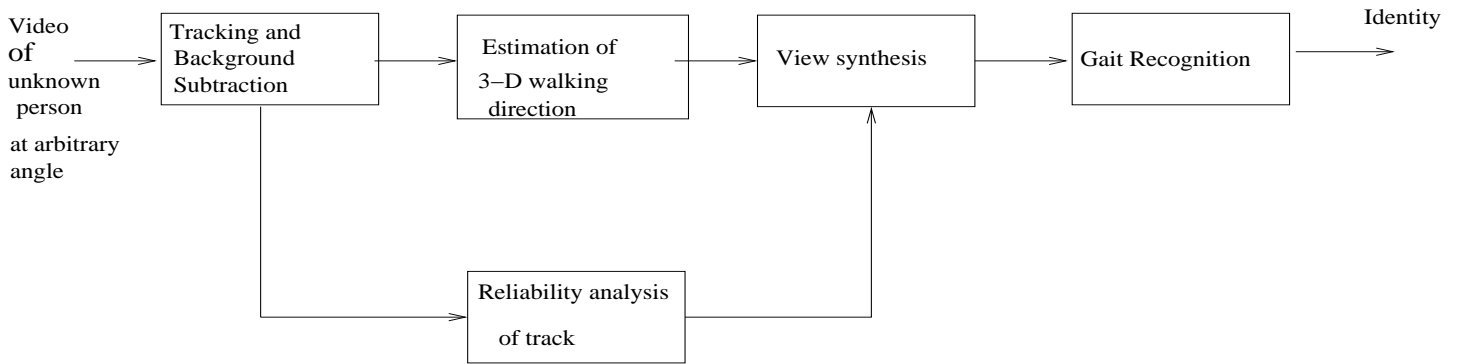


Figure 3: Framework for View Invariant Gait Recognition

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 (Figure 4). By tracking the direction of motion, α , in the video sequence, one can estimate the 3-D angle θ . This can be done by using the optical flow based structure from motion (SfM) equations. Under the assumption of planarity, 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. We refer to this approach as the “implicit SfM” approach. In the case where there is no real translation of the person e.g. person walking on a treadmill, an alternative approach is employed to obtain the synthesized views of the person. Given a set of point correspondences for a planar surface between the canonical and non-canonical views in a set of training images, we compute a homography. This homography is then applied to the binary silhouette of the person to obtain the synthesized views. We refer to this approach as the “homography approach”.

An overview of our gait recognition framework is given in Figure 3. We have reported recognition experiments [34] using two publicly available gait databases (NIST and CMU). The implicit SfM approach is used for the NIST databases while the homography approach is used for the CMU database. Keeping in view the limited quantity of training data, the DTW algorithm [35] is used for gait recognition. Acceptable recognition results are obtained for θ less than 30° . A by-product of the above method is a simple algorithm to synthesize novel views of a planar scene.

5 Human Recognition Using Face

Though face recognition has been intensively investigated for more than ten years, the state-of-the-art face recognition systems yield satisfactory performance only when confronted with controlled conditions. Unconstrained conditions such as illumination/pose variations and surveillance video scenarios impose significant challenges to existing recognition systems. Below, we summarize emerging techniques dealing with face recognition under illumination/pose variations and from videos. Since different approaches experimented on different datasets, comparison of recognition performance is not appropriate and hence no performance is actually reported below.

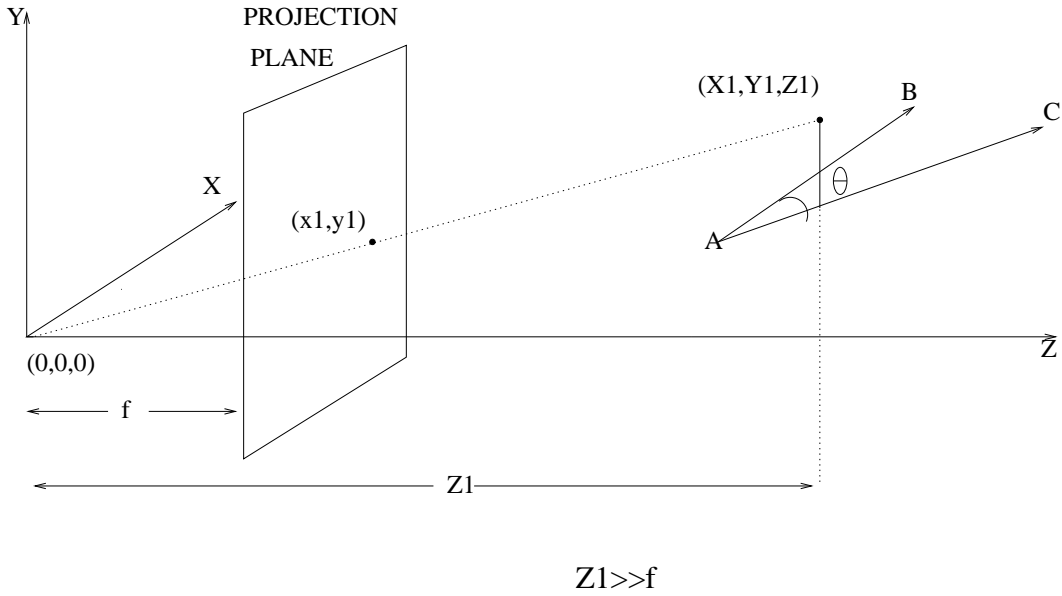


Figure 4: Imaging Geometry

5.1 Face Recognition under Illumination Variation

Recent works on face recognition under illumination variations [3, 23, 66, 84] employ a Lambertian reflectance model with a varying albedo field. A pixel h^s under a distant illuminant s is formulated as

$$h^s = p\mathbf{n}^T\mathbf{s} = \mathbf{t}^T\mathbf{s}; \quad \mathbf{n}_{3\times 1} \doteq [n_x, n_y, n_z]^T; \quad \mathbf{t}_{3\times 1} \doteq p\mathbf{n}, \quad (13)$$

where p is the albedo at the pixel, \mathbf{n} is the unit surface normal vector at the pixel, and \mathbf{s} (a 3×1 unit vector multiplied by its intensity) specifies the distant illuminant s . For an image \mathbf{h}^s , a collection of d pixels $\{h_i^s, i = 1, \dots, d\}$, by stacking all the pixels into a column vector, we have

$$\mathbf{h}^s \doteq [h_1^s, h_2^s, \dots, h_d^s]^T = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_d]^T \mathbf{s} = \mathbf{T} \mathbf{s}, \quad (14)$$

where $\mathbf{T} \doteq [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_d]^T$ contains complete albedo and shape information for the object and is called as *object-specific albedo-shape* matrix [84]. Therefore, the Lambertian reflectance model, when the attached and cast shadows are ignored, implies a rank-3 subspace [65] in which the appearances under different illumination reside.

If attached shadows are considered [3], the rank grows to infinity, but the energy is largely packed in a few harmonics components, thereby enabling a low-dimensional subspace approximation. However, in [3, 23], for one object to be recognized, one must have multiples (≥ 3) images stored in a gallery set, which is very inconvenient in practice. Essentially, generalization across illumination variation is offered by the illumination model, but no generalization between identities is available.

The requirement of storing multiple images is relaxed in [66, 84]: Only the training set stores multiple observations for multiple objects and the gallery set stores only one image per object. Here, a continuous-valued identity signature is used and a linear generalization from the training set to the gallery/probe set is assumed. The difference between [66] and [84] lies in how the linear blending coefficients are learnt. Once learnt, the blending coefficients offer an illumination-invariant signature of the identity. Both approaches can recognize probe images under illumination different from those of gallery images. The quotient image approach in [66] assumes that the

shapes of all objects are same and the albedo field of an unknown object lie in the rational span of the training set. The approach in [84] poses a rank constraint on the product of the albedo and surface normal. Below, we briefly review the approach in [84].

It states that any \mathbf{T} matrix can be represented as a linear combination of some basis matrices $\{\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_m\}$ coming from some m hypothetical base objects. Mathematically, there exist coefficients f_i 's such that

$$\mathbf{T} = \sum_{i=1}^m f_i \mathbf{T}_i = [\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_m](\mathbf{f} \otimes \mathbf{I}_3) = \mathbf{W}(\mathbf{f} \otimes \mathbf{I}_3), \quad (15)$$

where $\mathbf{f}_{m \times 1} \doteq [f_1, f_2, \dots, f_m]^T$, $\mathbf{W}_{d \times 3m} \doteq [\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_m]$, and \otimes denotes the matrix Kronecker (tensor) product. So, an image \mathbf{h}^s can be re-expressed as

$$\mathbf{h}^s = \mathbf{T}\mathbf{s} = \mathbf{W}(\mathbf{f} \otimes \mathbf{I}_3)\mathbf{s} = \mathbf{W}(\mathbf{f} \otimes \mathbf{s}). \quad (16)$$

Since the coefficient vector \mathbf{f} only relates the albedos and surface normals of the basis matrices, it has no relationship with the illumination \mathbf{s} . Thus, \mathbf{f} is an illumination-invariant description of the identity and is an appropriate quantity for face recognition under illumination variation. Eq. (16) presents a bilinear relationship between \mathbf{f} and \mathbf{s} . Once the \mathbf{W} matrix is given, the \mathbf{f} vector can be easily recovered using a bilinear algorithm.

However, learning the \mathbf{W} matrix from a training set of images is not a trivial task. In [20], the recovered \mathbf{W} minimizes the approximation error in the mean square sense and need not satisfy the integrability constraint. In other words, the hypothetical base objects in \mathbf{W} are not integrable. In [84], the recovered \mathbf{W} minimizes the above approximation error as well as a cost function that enforces the integrability constraint. As a consequence, [20] can only process the image ensemble consisting different objects under the same set of lighting sources (e.g. the case considered here) while [84] can process the image ensemble consisting of different objects under completely different lighting condition. Figure 5 shows the recovered \mathbf{W} using the algorithm developed in [84].

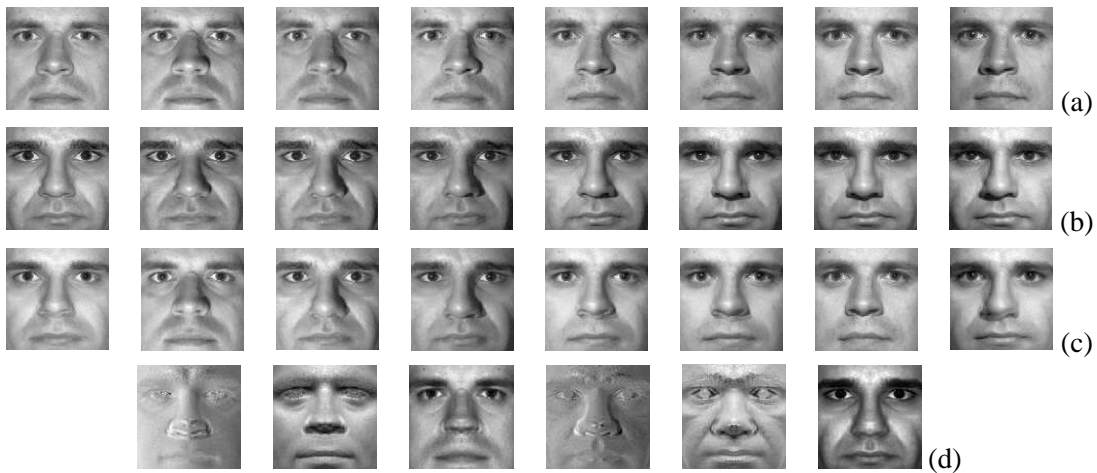


Figure 5: (a) The first basis object under eight different illuminations. (b) The second basis object under the same set of eight different illuminations. (c) Eight images (constructed by random linear combinations of two basis objects) illuminated by eight different lighting sources. (d) Recovered class-specific albedo-shape matrix \mathbf{W} showing the product of varying albedos and surface normals of two basis objects (i.e. the three columns of \mathbf{T}_1 and \mathbf{T}_2) using the algorithm in [84].

5.2 Face Recognition under Pose Variation

The issue of pose essentially amounts to a correspondence problem. If dense correspondences across poses are available and if a Lambertian reflectance model is further assumed, a rank-1 constraint is implied because theoretically, a 3D model can be recovered and used to render novel poses. However, recovering a 3D model from 2D images is a difficult task. There are two types of approaches: model-based and image-based. Model-based approaches [21, 64, 60, 6] require explicit knowledge of prior 3D models, while image-based approaches [57, 39, 47, 43] do not use prior 3D models. In general, model-based approaches [21, 64, 60, 6] register the 2D face image to 3D models that are given beforehand. In [21, 64], a generative face model is deformed through bundle adjustment to fit 2D images. In [60], a generative face model is used to regularize the 3D model recovered using the SfM algorithm. In [6], 3D morphable models are constructed based on many prior 3D models. There are mainly three types of image-based approaches: Structure from motion (SfM) [57], visual hull [39, 47], and light field rendering [43, 24] methods. The SfM approach [57] using sparse correspondence does not reliably recover the 3D model amenable for practical use. Recently developed methods [61] using optical flow and FFT computation show promise. The visual hull methods [39, 47] assume that the shape of the object is convex, which is not always satisfied by the human face, and also require accurate calibration information. The light field rendering methods [43, 24] relax the requirement of calibration by a fine quantization of the pose space and recover a novel view by sampling the captured data that form the so-called light field. Figure 6 illustrates the concept using a simple example of the 2D light-field of a 2D object.

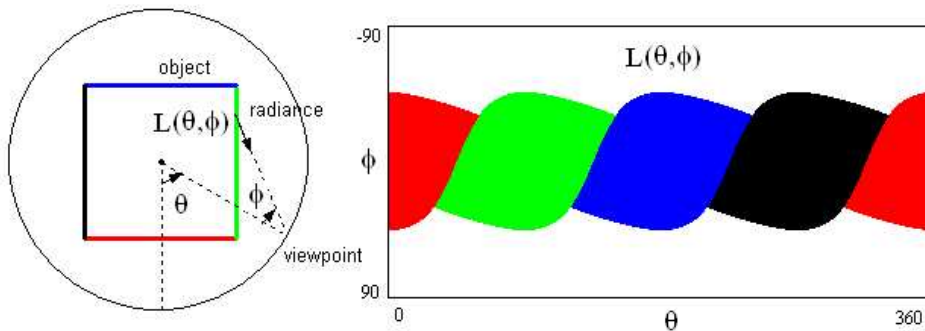


Figure 6: This figure illustrates the 2D light-field of a 2D object (a square with four differently colored sides), which is placed within a circle. The angles θ and ϕ are used to relate the viewpoint with the radiance from the object. The right image shows the actual light field for the square object. See another illustration in [26].

As mentioned earlier, pose variation essentially amounts to a correspondence problem. If dense correspondences across poses are available and a Lambertian reflectance is assumed, then a rank-1 constraint is implied. Unfortunately, finding correspondences is a very difficult task and, therefore there exist no subspace based on an appearance representation when confronted with pose variation. Approaches to face recognition under pose variation [54, 23, 26] avoid the correspondence problem by sampling the continuous pose space into a set of poses, *v.i.z.* storing multiple images at different poses for each person at least in the training set. In [54], view-based ‘Eigenfaces’ are learned from the training set and used for recognition. In [23], a denser sampling is used to cover the pose space. However, [23] uses object-specific images and appearances belong to a novel object (*i.e.* not in the training set) cannot be handled. In [26], the concept of light field [43] is used to characterize the continuous pose space. ‘Eigen’ light fields are learnt from the training set. However, the implementation of [26] still discretizes

the pose space and recognition can be based on probe images at poses in the discretized set. One should note that the light field is not related to variation in illumination.

5.3 Face Recognition under Illumination and Pose Variations

Approaches to handling both illumination and pose variations include [6, 25, 74, 83, 82]. The approach [6] uses morphable 3D models to characterize the human faces. Both geometry and texture are linearly spanned by those of the training ensemble consisting of 3D prior models. It is able to handle both illumination and pose variations. Its only weakness is a complicated fitting algorithm. Recently, a fitting algorithm more efficient than [6] is proposed in [59]. In [25], the Fisher light field is proposed to handle both illumination and pose variations, where the light field is used to cover the pose variation and the Fisher discriminant analysis to cover the illumination variation. Since discriminant analysis is a statistical analysis tool which minimizes the within-class scatter while maximizing between-class scatter and has no relationship with any physical illumination model, it is questionable that discriminant analysis is able to generalize to new lighting conditions. Instead, this generalization may be inferior because discriminant analysis tends to overly tune to the lighting conditions in the training set. The ‘Tensorface’ approach [74] uses a multilinear analysis to handle various factors such as identity, illumination, pose, and expression. The factors of identity and illumination are suitable for linear analysis, as evidenced by the ‘Eigenface’ approach (assuming a fixed illumination and a fixed pose) and the subspace induced by the Lambertian model, respectively. However, the factor of expression is arguably amenable for linear analysis and the factor of pose is not amenable for linear analysis. In [83], preliminary results are reported by first warping the albedo and surface normal fields at the desired pose and then carrying on recognition as usual. The approach in [82] extends the algorithm in [84] that is for illumination variation to deal with pose variation as well. In [84], all face images are in frontal view and hence no pose variances are present. In [82], we consider a finite set of views that cover from the left profile to the right profile almost uniformly. By treating the images consisting of all views of the same individual illuminated by the same lighting source as an ‘augmented’ image, we can apply the algorithm developed in [84] to the ‘augmented’ image since it is illuminated by one source. Figure 7 shows the part of the \mathbf{W} matrix learned using images in the PIE database [67].

5.4 Face Recognition from Videos

Various approaches for performing face recognition based on video sequences have been proposed. However, most approaches [9] are essentially still-image-based and treat each video frame separately. Typically, they first perform face tracking and then perform recognition based on one or several tracked face regions that satisfy certain criteria. In the above strategy, two important characteristics of a video sequence are disregarded.

- Multiple looks. A video sequence provides a large amount of observations. This is very attractive considering the projective nature of the imaging geometry and the uncontrolled lighting distribution. All these factors, coupled with personal variations such as facial expression, make the 2D face appearances of one individual possess infinitely many possibilities. However, the above typical approach makes recognition decision based only on a sparse set of observations.
- Temporal continuity. Frames of a video sequence come in a sequential fashion and possess certain smoothness between successive frames that is referred as temporal continuity. Such continuity is often exploited in developing tracking algorithms. However, there is psychophysical evidence [37] suggesting that the temporal continuity is also important for recognition.



Figure 7: The first 9 columns of the learned \mathbf{W} matrix.

We now highlight some recent approaches to face recognition from videos, along the line of utilizing the video characteristics.

In [79], two mutual view subspaces are pre-learned. In testing, for each frame of a video sequence, the algorithm computes its similarity score with both subspaces and takes the higher ones. A time-evolving curve of similarity score is plotted. If the temporal information is stripped, this method is similar to view-based and modular eigenspaces proposed by Pentland et al. [54]. Evidence integration is also performed in [17, 44]. Both [17, 44] are based on the framework of active appearance model [17]. In [44], the kernel discriminant analysis features are extracted from the image warped to a frontal view for recognition.

Another line of research effort summarizes the appearances presented in the video sequence. Examples of such a method are [63, 19, 78]. In [63], a multivariate Gaussian density is fitted for a video sequence, where all facial images cropped out by a separate tracker are assumed to be i.i.d. samples from that distribution. Recognition is performed by comparing the Kullback-Leibler divergence distance [11] between the gallery and probe videos. However, a multivariate Gaussian density is deemed to have difficulty in modeling significant variations caused by pose and illumination. In [19], manifolds are formed for multiple images. Recognition is performed by computing the shortest distance between two manifolds. The manifold takes a certain parameterized form and the parameters

are directly learned from the visual appearances. In [78], principal subspaces are learned for multiple images and principal angle between two principal subspaces is used for recognition. The computation of principal angle is also carried on the feature space embedded by kernel functions. One common disadvantage of the above two approaches is that they also assume that the face regions already been cropped beforehand, coming from either a detector or a tracker.

However, when the above three approaches utilize the multiple appearances provided by the video sequence, they do not take into account the temporal continuity between successive video frames. We now show approaches that utilize the temporal coherence embedded in the video sequences.

In [85], still-to-video recognition is solved, where the gallery consists of still images and the probes are video sequences. Since the detected face might be moving in the video sequence, we inevitably have to deal with uncertainty in tracking as well as in recognition. Rather than resolving these two uncertainties separately, our strategy is to perform simultaneous tracking and recognition of human faces from a video sequence. A time series state space model is proposed to fuse temporal information in a probe video, which simultaneously characterizes the kinematics and identity using a motion vector and an identity variable, respectively. The joint posterior distribution of the motion vector and the identity variable is estimated at each time instant and then propagated to the next time instant. Marginalization over the motion vector yields a robust estimate of the posterior distribution of the identity variable. A computationally efficient sequential importance sampling (SIS) algorithm [16] is used to estimate the posterior distribution. Empirical results demonstrate that, due to the propagation of the identity variable over time, a degeneracy in posterior probability of the identity variable is achieved.

In addition, the gallery can be extend to have videos as inputs. A learning algorithm to automatically extract exemplars from the gallery video sequences has been described in [85]. To represent each object in the gallery, multiple exemplars are extracted. During test, the SIS is adopted to use the temporal coherence to boost recognition performance.

In [46], hidden Markov models are used to learn the dynamics before successive appearances. Matching video sequences is equivalent to comparing two Markov models. In [40], pose variations are handled by learning the view-discretized appearance manifolds from the training ensemble. Transition probabilities from one view to another view are used to regularize the search space. However, in [46, 40], the cropped images are used for testing. Recently, linear dynamical system model [68] has been used to model the video sequence and the system model coefficients are used in face recognition [1].

6 Conclusions

In this chapter, we have briefly described some of the methods that we have developed towards the goal of human recognition using biometrics such as face and gait. While face recognition has been an area of research in computer vision for many years, the use of gait for recognition is a more recent phenomenon. Obviously, there are a number of limitations in the use of these techniques in uncontrolled environments, e.g outdoors under various lighting conditions. Progress in computer vision research and allied fields like image/video processing, pattern recognition, and machine learning will allow us to develop more realistic algorithms in the future. Most of the present face recognition algorithms work with a frontal view of the face, while most gait recognition methods assume a side view of the person. The full potential of face recognition using video sequences still needs to be explored and its advantage vis-a-vis still images needs to be studied. The problems that affect the performance of both these techniques are the effects of variable illumination, time elapsed and pose invariance. Today, multimodal biometrics are becoming more popular in order to achieve high recognition rates. An example of fusing face and

gait signatures may be found in [36]. The idea in multimodal biometrics is to combine different cues like face, gait, fingerprint, iris, ear in order to develop an identifying signature of the individual. Efficient means for combining some or all of these different biometrics automatically is still an open question.

References

- [1] G. Aggarwal, A. Roy-Chowdhury, and R. Chellappa. A system identification approach for video-based face recognition. *International Conference on Pattern Recognition*, Cambridge, UK, August 2004.
- [2] C.D. Barclay, J. E. Cutting, and L.T. Kozlowski. Temporal and spatial factors in gait perception that influence gender recognition. *Perception and Psychophysics*, 23:145–152, 1978.
- [3] R. Basri and D.W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(2):218–233, February 2003.
- [4] T. Beardsworth and T. Buckner. The ability to recognize oneself from a video recording of ones movements without seeing ones body. *Bulletin of the Psychonomic Society*, 18(1):19–22, 1981.
- [5] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.
- [6] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, September 2003.
- [7] A.F. Bobick and A. Johnson. Gait recognition using static activity-specific parameters. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, December 2001.
- [8] 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*, pages 267–272, Washington D.C., May 2002.
- [9] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland. Multimodal person recognition using unconstrained audio and video. In *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pages 176–181, Washington, 1999.
- [10] R. Collins, R. Gross, and J. Shi. Silhouette-based human identification from body shape and gait. *Proceedings of IEEE Conference on Face and Gesture Recognition*, May 2002.
- [11] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley, 1991.
- [12] 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*, pages 43–48, 1999.
- [13] J. Cutting and L. Kozlowski. Recognizing friends by their walk:gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, 9:353–356, 1977.
- [14] J. E. Cutting and D.R. Proffitt. *Gait perception as an example of how we perceive events*. Plenum Press, London, 1981.
- [15] W.H. Dittrich. Action categories and the perception of biological motion. *Perception*, 22:15–22, 1993.
- [16] A. Doucet, N. de Freitas, and N. (eds) Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.
- [17] G. Edwards, C. Taylor, and T. Cootes. Improving identification performance by integrating evidence from sequences. *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1999.
- [18] O.D. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, 1993.
- [19] A. Fitzgibbon and A. Zisserman. Joint manifold distance: a new approach to appearance based clustering. *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003.

- [20] W. T. Freeman and J. B. Tenenbaum. Learning bilinear models for two-factor problems in vision. *Prof. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997.
- [21] P. Fua. Regularized bundle adjustment to model heads from image sequences without calibrated data. *International Journal of Computer Vision*, 38:153–157, 2000.
- [22] Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust., Speech, Signal Processing*, 29(2):254–272, April 1981.
- [23] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.
- [24] S.J. Gortler, R. Grzeszczuk, R. Szeliski, and M. Cohen. The lumigraph. *Proceedings of SIGGRAPH*, pages 43–54, New Orleans, LA, USA, 1996.
- [25] R. Gross, I. Matthews, and S. Baker. Fisher light-fields for face recognition across pose and illumination. *Proc. of the 24th Symposium of the German Association for Pattern Recognition (DAGM)*, 2002.
- [26] R. Gross, I. Matthews, and S. Baker. Eigen light-fields and face recognition across pose. *Prof. Face and Gesture Recognition*, Washington D.C., May 2002.
- [27] 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.
- [28] G.F. Harris and P.A. Smith (Editors). *Human Motion Analysis: Current Applications and Future Directions*. IEEE Press, 1996.
- [29] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.
- [30] A. Howell and H. Buxton. Face recognition using radial basis function neural networks. *Proc. British Machine Vision Conference*, pages 455–464, 1996.
- [31] P.S. Huang, C.J. Harris, and M.S. Nixon. Recognizing humans by gait via parametric canonical space. *Artificial Intelligence in Engineering*, 13(4):359–366, October 1999.
- [32] G. Johansson. Visual perception of biological motion and a model for its analysis. *PandP*, 14(2 1973):201–211, 1973.
- [33] G. Johansson. Visual motion perception. *Scientific American*, 232:76–88, 1975.
- [34] A. Kale, A.K. Roy Chowdhury, and R. Chellappa. Towards a view invariant gait recognition algorithm. *Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 143–150, 2003.
- [35] A. Kale, N. Cuntoor, B. Yegnanarayana, A.N. Rajagopalan, and R. Chellappa. Gait analysis for human identification. In *Proceedings of Audio, Video and Biometric Person Authentication*, 2003.
- [36] A. Kale, A. Roy-Chowdhury, and R. Chellappa. Fusion of gait and face for human identification. *Intl. Conf. on Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004.
- [37] B. Knight and P. Johnston. The role of movement in face recognition. *Visual Cognition*, 4:265–274, 1997.
- [38] L. Kozlowski and J. Cutting. Recognizing the sex of a walker from a dynamic point display. *Perception and Psychophysics*, 21:575–580, 1977.
- [39] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994.
- [40] K. Lee, J. Ho, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003.
- [41] L. Lee and G. Dalley. Learning pedestrian models for silhouette refinement. In *International Conf. on Computer Vision*, Nice, France, 2003.

- [42] L. Lee and W.E.L. Grimson. Gait analysis for recognition and classification. *Proceedings of the IEEE Conference on Face and Gesture Recognition*, pages 155–161, 2002.
- [43] M. Levoy and P. Hanrahan. Light field rendering. *Proc. SIGGRAPH*, 1996.
- [44] Y. Li, S. Gong, and H. Liddell. Constructing facial identity surfaces for recognition. *International Journal of Computer Vision*, 53(1):71–92, 2003.
- [45] J. Little and J. Boyd. Recognizing people by their gait: the shape of motion. *Videre*, 1(2):1–32, 1998.
- [46] X. Liu and T. Chen. Video-based face recognition using adaptive hidden Markov models. *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003.
- [47] W. Matusik, C. Buehler, R. Raskar, and L. Gortler, S. and McMillan. Image-based visual hulls. *Proceedings of SIGGRAPH*, pages 369 – 374, 2000.
- [48] S. McKenna and S. Gong. Non-intrusive person authentication for access control by visual tracking and face recognition. *Proc. Intl. Conf. Audio- and Video-based Biometric Person Authentication*, pages 177–183, 1997.
- [49] B. Moghaddam. Principal manifolds and probabilistic subspaces for visual recognition. *IEEE Trans. PAMI*, 24(6):780–788, 2002.
- [50] B. Moghaddam and A.P. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19:696–710, July 1997.
- [51] M.P. Murray, A.B. Drought, and R.C. Kory. Walking patterns of normal men. *Journal of Bone and Joint surgery*, 46-A(2):335–360, 1964.
- [52] E. Muybridge. *The Human Figure in Motion*. Dover Publications, 1901.
- [53] P. Penev and J. Atick. Local feature analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, 7:477–500, 1996.
- [54] A.P. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1994.
- [55] 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.
- [56] 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.
- [57] G. Qian and R. Chellappa. Structure from motion using sequential monte carlo methods. *Proceedings of IEEE International Conference on Computer Vision*, 2:614–621, Vancouver, Canada, 2001.
- [58] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, February 1989.
- [59] S. Romdhani and T. Vetter. Efficient, robust and accurate fitting of a 3d morphable model. *Proceedings of IEEE International Conference on Computer Vision*, pages 59–66, Nice, France, 2003.
- [60] A. Roy Chowdhury and R. Chellappa. Face reconstruction from video using uncertainty analysis and a generic model. *Computer Vision and Image Understanding*, 91:188–213, 2003.
- [61] A. Roy Chowdhury and R. Chellappa. Stochastic approximation and rate distortion analysis for robust structure and motion estimation. *International Journal of Computer Vision*, 55(1):27–53, 2003.
- [62] W. I Scholhorn, . Nigg B.M, D.J.Stephanshyn, and W. Liu. Identification of individual walking patterns using time discrete and time continuous data sets. *Gait and Posture*, 15:180–186, 2002.

- [63] G. Shakhnarovich, J. Fisher, and T. Darrell. Face recognition from long-term observations. *European Conference on Computer Vision*, Copenhagen, Denmark, May 2002.
- [64] Y. Shan, Z. Liu, and Z. Zhang. Model-based bundle adjustment with application to face modeling. *Proceedings of International Conference on Computer Vision*, pages 645–651, Vancouver, Canada, 2001.
- [65] A. Shashua. On photometric issues in 3D visual recognition from a single 2D image. *International Journal of Computer Vision*, 21:99–122, 1997.
- [66] A. Shashua and T. R. Raviv. The quotient image: Class based re-rendering and recognition with varying illuminations. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23:129–139, 2001.
- [67] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. *Prof. Face and Gesture Recognition*, 2002.
- [68] S. Soatto, G. Doretto, and Y.N. Wu. Dynamic textures. *International Conference on Computer Vision*, 2001.
- [69] S. V. Stevenage, M. S. Nixon, and K. Vince. Visual analysis of gait as a cue to identity. *Applied Cognitive Psychology*, (13):513–526, March 1999.
- [70] A. Sundaresan, A. Roy Chodhury, and R. Chellappa. A hidden markov model based framework for recognition of humans from gait sequences. *Proceedings of the International Conference Image Processing*, September 2003.
- [71] R. Tanawongsuwan and A.F. Bobick. Gait recognition from time-normalized joint-angle trajectories in the walking plane. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages II:726–731, 2001.
- [72] D. Tolliver and R. Collins. Gait shape estimation for identification. *Proceedings of AVBPA*, pages 734–742, 2003.
- [73] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:72–86, 1991.
- [74] M.A.O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. *European Conference on Computer Vision*, 2350:447–460, Copenhagen, Denmark, May 2002.
- [75] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa. Role of shape and kinematics in human movement analysis. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Washington D.C., 2004.
- [76] H. Wechsler, V. Kakkad, J. Huang, S. Gutta, and V. Chen. Automatic video-based person authentication using the RBF network. *Proc. Intl. Conf. on Audio- and Video-based Biometric Person Authentication*, pages 85–92, 1997.
- [77] L. Wiskott, J. M. Fellous, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19:775–779, 1997.
- [78] L. Wolf and A. Shashua. Kernel principal angles for classification machines with applications to image sequence interpretation. *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, WI, USA, 2003.
- [79] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. *Proc. Intl. Conf. on Automatic Face and Gesture Recognition*, Nara, Japan, October, 1998.
- [80] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. *Proceedings, International Conference on Automatic Face and Gesture Recognition*, pages 336–341, 1998.
- [81] W. Zhao, R. Chellappa, A. Rosenfeld, and J. Phillips. Face recognition: A literature survey. *ACM Computing Surveys*, 12:399 – 458, 2003.
- [82] S. Zhou and R. Chellappa. Illuminating light field: Image-based face recognition across illuminations and poses. *IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, Korea, May 2004.
- [83] S. Zhou and R. Chellappa. Rank constrained recognition under unknown illumination. *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, Nice, France, 2003.

- [84] S. Zhou, R. Chellappa, and D. Jacobs. Characterization of human faces under illumination variations using rank, integrability, and symmetry constraints. *European Conference on Computer Vision*, Prague, The Czech Republic, May 2004.
- [85] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91:214–245, July-August 2003.