Using Bilinear Models for View-invariant Identity Recognition from Gait

Fabio Cuzzolin and Stefano Soatto

Abstract

Human identification from gait is a challenging task in realistic surveillance scenarios in which people walking along arbitrary directions are shot by a single camera. In this paper we address the problem of finding an effective view-invariant approach for the human ID from gait problem. Given a dataset of sequences in which different people walking normally are seen from several well-separated views, we learn bilinear models to classify the identity of a known person from a view not included in the available training set. Assuming fixed dynamics, each sequence can be mapped to an observation vector with a constant number of poses using a Markov model. We test our approach on the CMU Mobo database, showing how even adopting a rather compact representation for images, bilinear separation of ID and view outperforms the standard baseline algorithm.

1. Introduction

Biometrics has been receiving a growing attention in the last decade, as automated identification systems became essential in the context of surveillance and security. In addition to standard biometrics like face recognition of fingerprint comparison, people have started to work on non-cooperative approaches in which the person to identify moves freely in the surveyed environment, and is possibly unaware of his/her identity being checked. In this perspective, the problem of recognizing people from natural gait has been studied by several people as a non-intrusive biometric approach [36], starting from a seminal work of Niyogi and Adelson [26].

A variety of techniques have been introduced, most based on silhouette analysis [5, 37]. Many other gait signatures, however, have been studied, ranging from optical flow [22], to velocity moments [30] shape symmetry [14], frieze patterns [23], height and stride estimation [2], static body parameters [16], or area-based metrics [11]. Sensor fusion approaches to combine evidence for identity recognition have also been proposed [4, 29, 8]. Concerning classification, a number of methods apply some pattern recognition technique after dimensionality reduction (through eigenspaces [1, 25], or PCA/MDA [34, 13]). Others employ stochastic models like HMMs to describe the dynamics of the gait [17, 15].

In the last few years, the field has been somehow organized in order to allow the comparison of the different approaches. This has led to the creation of a number of public databases, which can be used as a common ground on which validate the algorithms. One of the most popular is perhaps the USF database [28], in which the problem is posed in a realistic, outdoor context with cameras located at a distance, and people walking on an elliptical path. The experiments are designed to study the effect of many factors (covariates) on the identity classification, like time, clothing, ground, shoes, and view. However, the experiments contemplate only two cameras at fairly close viewpoints (some 30 degrees), while people is shot while walking along the opposite side of the ellipse, so that the resulting views are almost fronto-parallel. As a matter of fact, appearance-based algorithm comparing the appearance of silhouettes seem to work well in the experiments concerning viewpoint variability, while one would guess they should not for widely separated views. In a realistic setup, in fact, it is reasonable to assume that the person to identify would walk in the surveyed area from an arbitrary direction. View-invariance is then a crucial issue to make identification from gait suitable for real-world applications. This matter has been recently investigated by several research groups [35, 38, 3, 18, 29, 16].

In this paper we focus on the issue of building an effective view-independent algorithm for identity recognition from gait by means of bilinear models [33]. A single camera system would acquire in time a dataset of sequences in which people are seen walking from several viewpoints. As they are capable to learn how view and identity interact in such a mixed training set, bilinear models allow to build a classifier which, given a new sequence in which a known person is seen from a view not in the training set, can iteratively estimate both identity and view parameters, significantly improving recognition performances.

We propose a preprocessing stage in which each sequence is given as input to a hidden Markov model. Assuming fixed dynamics, the HMM would cluster it into a fixed number of poses, no matter the speed or starting position, producing an

observation vector that can then be fed to a bilinear model. In a sense, this correspond to building a meta-model with two different layers. We use the silhouettes provided by the CMU Mobo database [12] to test our approach, showing how even adopting a rather compact representation for images, bilinear separation of ID and view outperforms the standard baseline algorithm.

2. A realistic scenario

In real-world applications identity recognition from gait has to cope with challenging surveillance scenarios in which a single camera controls a large room in which people walk at a distance, coming from arbitrary directions (for instance an airport, or a hotel hall). An equivalent description of the problem can be given in terms of view invariance: a person's gait should be recognized no matter the viewpoint from which it is seen. The view-invariance issue in the gait ID context has been actually studied by many people. This of course can be addressed in several ways.

If a 3D articulated model of the moving person is available, tracking can be used as a preprocessing stage to drive recognition. Cunado et al. [6], for instance, used their evidence gathering technique to analyze the leg motion in both walking and running gait, using two different models employing coupled oscillators and the biomechanics of human locomotion as underlying concepts. They provide estimates of the inclination of thigh and leg, by deriving a phase-weighted Fourier description gait signature in an automatic way. Yam et al [38] also worked on a similar model-based approach. Urtasun and Fua [35] proposed an approach to gait analysis that relies on fitting 3D temporal motion models to synchronized video sequences. These models allowed them to recover motion parameters that can be used to recognize people, in particular the coefficients of the singular value decomposition of the estimated model angles. Bhanu and Han [3] also matched a 3D kinematic model to 2D silhouettes, extracting a number of feature angles from the fitted model, and using them for gait recognition.

It is rather well known from the literature, though, that model-based 3D tracking is a difficult problem, as manual initialization is often required, and optimization in a high-dimensional parameter space is sensitive to convergence defects. Kale et al. [18] have instead proposed a method to generate a synthetic side-view of the moving person using a single camera, if the person is far enough. They use two different methods, one based on the perspective projection model, and the other applying structure from motion to the optical flow. Shakhnarovich et al. [29] suggested a view-normalization approach in a multiple camera context. They used the volumetric intersection of the visual hulls of all the camera silhouettes to get a volumetric reconstruction of the moving body, and a trajectory analysis to estimate the virtual side view of the obtained voxelset (plus a frontal view for integrated face recognition). Johnson e Bobick also presented a multi-view gait recognition method using static body parameters recovered during the walking motion across multiple views [16].

Here we focus on the single camera scenario, and adopt bilinear models to build a system able to recognize people from their gait no matter the viewpoint from which they are shot. We will show how to use a Markov model to preprocess each sequence in order to cluster it into a fixed number of poses (under certain assumptions on the dynamics of the motion). Given a training set of motions characterized by different directions (or equivalently shot by different viewpoints) we feed the resulting vectors to an asymmetric bilinear model, and use it to classify the new sequences in which a known person walks in a different direction (viewpoint).

3. Bilinear models

Bilinear models have been introduced by Tenenbaum et al. [33] as a tool for separating what they call "style" and "content" of objects to classify, meaning two distinct class labels of the same objects. Common but useful examples can be font and letters in writing, or word and accent in speaking.

In the *symmetric* model, style and content are represented by two parameter vectors $\mathbf{a}^s$ and $\mathbf{b}^c$ with dimension I and J respectively. Given a training set of K-dimensional observations $\{\mathbf{y}_k^{sc}\}$, $k = 1, .., K$ with two different labels $s \in [1, .., S]$ (style) and $c \in [1, .., C]$ (content), we assume it can be described by a bilinear model of the type

$$\mathbf{y}_k^{sc} = \sum_{i=1}^{I} \sum_{j=1}^{J} w_{ijk} a_i^s b_j^c \tag{1}$$

that, letting $\mathbf{W}_k$ denote the k-th matrix of dimension $I \times J$ with entries w_{ijk}, can be rewritten as

$$\mathbf{y}_k^{sc} = (\mathbf{a}^s)^T \mathbf{W}_k \mathbf{b}^c.$$

The matrices $\mathbf{W}_k$ define a *bilinear map* from the style and content spaces to the K-dimensional observation space.

When the interaction factors can vary with style (i.e. w^s_{ijk} depend on s) we get an *asymmetric* model

$$\mathbf{y}^{sc} = \mathbf{A}^s \mathbf{b}^c \tag{2}$$

where $\mathbf{A}^s$ denotes the $K \times J$ matrix with entries $\{a^s_{jk} = \sum_i w^s_{ijk} a^s_i\}$, a style-specific linear map from the content space to the observation space.

3.1. Training an asymmetric model

Given a training set of observations with two labels, a bilinear model can be fitted to the data by means of simple linear algebraic techniques. If the training set has (roughly) the same number of measurements for each style and each content class, an asymmetric model can be fit to the data by singular value decomposition (SVD). Once stacked the training data into the $(SK) \times C$ matrix

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}^{11} & \cdots & \mathbf{y}^{1C} \\ \cdots & \cdots & \cdots \\ \mathbf{y}^{S1} & \cdots & \mathbf{y}^{SC} \end{bmatrix} \tag{3}$$

the asymmetric model can be written as $\mathbf{Y} = \mathbf{AB}$ where $\mathbf{A}$ and $\mathbf{B}$ are the stacked style and content parameter matrices,

$$\mathbf{A} = [\mathbf{A}^1 \cdots \mathbf{A}^S]', \quad \mathbf{B} = [\mathbf{b}^1 \cdots \mathbf{b}^C].$$

The least-square optimal style and content parameters are then easily found by computing the SVD of (3) $\mathbf{Y} = \mathbf{USV}^T$, and assigning

$$\mathbf{A} = [\mathbf{US}]_{col=1..J} \quad \mathbf{B} = [\mathbf{V}^T]_{row=1..J}.$$

If the training data are not equally distributed among the classes, the least-square optimum has to be found [33].

3.2. Content classification of unknown style

Suppose that we have learned a bilinear model from a training set of data, and a new set of observations becomes available in a new style, different from all those present in the training set, but with content labels between those learned in advance. In this case an iterative procedure can be set up to factor out the effects of style and classify the content labels of the test observations. As a matter of fact, knowing the content class assignments of the new data is easy to find the parameters for the new style $\tilde{s}$ by solving for $\mathbf{A}^{\tilde{s}}$ in the asymmetric model (2). Analogously, having a map $\mathbf{A}^{\tilde{s}}$ for the new style we could easily classify the test vectors by measuring their distance from $\mathbf{A}^{\tilde{s}}\mathbf{b}^c$ for each (known) content vector $\mathbf{b}^c$.

The question can be approached by fitting a mixture model to the learned bilinear model by means of the EM algorithm [9]. The EM algorithm alternates between computing the probabilities $p(c|\tilde{s})$ of the current content label given an estimate of the style (E step), and estimating the linear map for the unknown style given the current content class probabilities (M step). More precisely, we assume that the probability generated by the new style $\tilde{s}$ and content c is given by a Gaussian distribution

$$p(\mathbf{y}|\tilde{s}, c) \propto \exp - \frac{\|\mathbf{y} - \mathbf{A}^{\tilde{s}}\mathbf{b}^c\|^2}{2\sigma^2} \tag{4}$$

while its total probability[1] is $p(\mathbf{y}) = \sum_c p(\mathbf{y}|\tilde{s}, c)p(\tilde{s}, c)$ where in absence of prior information $p(\tilde{s}, c)$ is supposed to be equally distributed.

In the E step the algorithm calculates the probabilities

$$p(\tilde{s}, c|\mathbf{y}) = \frac{p(\mathbf{y}|\tilde{s}, c)p(\tilde{s}, c)}{p(\mathbf{y})}$$

and classifies the test data by finding the content class c which maximizes $p(c|\mathbf{y}) = p(\tilde{s}, c|\mathbf{y})$.

In the M step the style matrix which maximizes the log likelihood of the test data is estimated, yielding

$$\mathbf{A}^{\tilde{s}} = \frac{\sum_c \mathbf{m}^{\tilde{s}c}(\mathbf{b}^c)^T}{\sum_c n^{\tilde{s}c}\mathbf{b}^c(\mathbf{b}^c)^T}$$

[1]The general formulation allows the presence of more than one unknown style, [33].

where $\mathbf{m}^{\tilde{s}c} = \sum_{\mathbf{y}} p(\tilde{s}, c|\mathbf{y})\mathbf{y}$ is the mean observation weighted by the probability of having style $\tilde{s}$ and content c, and $n^{\tilde{s},c} = \sum_{\mathbf{y}} p(\tilde{s}c|\mathbf{y})$.

The effectiveness of the method depends critically on the goodness of the representation chosen for the observation vectors. A formal analysis of the applicability of this methodology is still needed, as a way of incorporating domain-specific knowledge in the algorithm.

3.3. Bilinear models ID-view

In the context of view-invariant identity recognition from gait, our observations (the sequences of silhouettes acquired through the cameras) are as a matter of fact dependent on several factors: in particular, identity (content) and viewpoint (style). It is then natural to formalize the problem of recognizing a known person from a new, unknown viewpoint as a classification problem in the bilinear context.

Elgammal and Lee have recently used them to separate pose and ID from a database of poses in the context of GaitID [20, 10]. Here we consider *each subsequence* as an observation, dependent on the two factors *identity* and *view*, and apply the technique of Section 3.2 to assign an identity to each test sequence.

4. Silhouette representation

A formal characterization of the desirable properties a training should meet to be well described by a bilinear model is still missing. However, they were originally presented as a way of finding *approximate* solutions to problems in which two factors are involved [33], without a precise knowledge of the behavior of the observation vectors in each specific task. In the gaitID context, Elgammal and Lee have analyzed the geometry of cycles in the visual space, and adopted local linear embedding [33] as a tool to re-sample homogeneously each cycle into a fixed number of poses.

Here instead, following the original aim of bilinear models, we adopt a different strategy, choosing a simple but effective feature representation of the silhouettes to reduce the computational load. We will see in the following how this does not affect the performance of the classification. In particular, given a silhouette we detect its center of mass, rescale it to the associated bounding box, and subdivide the resulting image into 5 "standard" regions by splitting the lower third in two halves, the second third in another two halves, and keeping the upper third whole (see Figure 1). When the view is frontal,

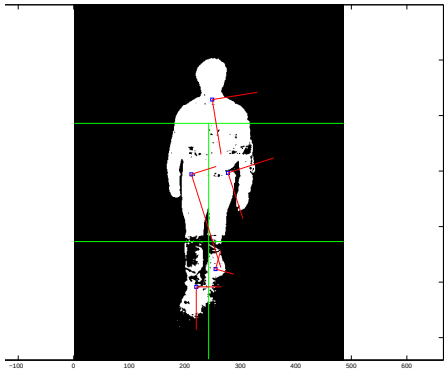

Figure 1: Feature extraction. The silhouette is divided into 5 regions (delimited by the green lines), the lengths of the principal axes (red lines) of the ellipsoid fitting each region are computed by SVD, and collected in a 10-dimensional vector.

those regions roughly correspond to head, arms and legs. For each region we compute the moments of the corresponding part of the silhouette by SVD, and collect the two largest eigenvalues (the lengths of the principal axes of the fitting ellipsoid). Similar approaches can be found in [29, 16, 21].

Each image is then associated with a 10-dimensional feature vector, two component for each region. All image sequences will then be encoded by a sequence of feature vectors, in general of different length. To make them suitable as input of a bilinear model learning stage (Section 3.1) we need to find a homogeneous representation. Hidden Markov models [24] provide us with a tool for transforming each sequence into a fixed-length observation vector.

5. Sequence re-sampling using HMMs

5.1. Hidden Markov modeling

A *hidden Markov model* is a statistical model whose states $\{X_k\}$ form a *Markov chain*; the only observable quantity is a corrupted version y_k of the state called *observation process*. Using the notation in [24] we can associate the elements of the finite state space $\mathcal{X} = \{1, ..., n\}$ with coordinate versors $e_i = (0, .., 0, 1, 0, .., 0) \in \mathbb{R}^n$ and write the model as

$$\begin{cases} X_{k+1} = AX_k + V_{k+1} \\ y_{k+1} = CX_k + diag(W_{k+1})\Sigma X_k \end{cases}$$

where $\{V_{k+1}\}$ is a sequence of martingale increments and $\{W_{k+1}\}$ is a sequence of i.i.d. Gaussian noises $\mathcal{N}(0, 1)$. The model parameters will then be the *transition matrix* $A = (a_{ij}) = P(X_{k+1} = e_i | X_k = e_j)$, the matrix C collecting the *means of the state-output distributions* (being the j-th column $C_j = E[p(y_{k+1}|X_k = e_j)]$) and the matrix Σ of the variances of the output distributions.

The set of parameters A, C and Σ of an HMM can be estimated, given a sequence of observations $\{y_1, ..., y_T\}$, through (again) an application of the Expectation-Maximization (EM) algorithm [24]. The likelihoods $\Gamma^i(y_{k+1})$ of the measurements y_{k+1} with respect to all the states $e_i, i = 1, ..., n$, are used to drive the recursive state estimation

$$\hat{X}_{k+1} = \sum_{i=1}^{n} A_i \langle \hat{X}_k, \Gamma^i(y_{k+1}) \rangle,$$

where n is the number of states, A_i is the i-th column of A and $\langle \cdot, \cdot \rangle$ is the usual scalar product.

5.2. Sampling and sequence representation

Given a sequence of feature vectors extracted from all the silhouettes of a sequence, EM yields as output a finite state representation of the motion, in which the transition matrix encodes the sequence's dynamics, while the columns of the C matrix are the poses representing each state in the observation space. There is no need to estimate the period of the cycle, as the poses are automatically associated with the states of the Markov model. Furthermore, sequences with variable speed cause no trouble, in opposition to methods based on the estimation of the fundamental frequency of the motion [22].

In the gait ID case, though, the dynamics is the same for all the sequences, being all of them instances of the walking motion, and can be factored out as the topology of the resulting HMM will be constant. If we also assume (as it is frequently done in the literature) that people are walking at constant speed, the transition matrix is of no use and can be neglected. If we apply the HMM-EM algorithm to all the input sequences with a *same* number of states, it provides a standardized representation in which each sequence is mapped to a constant number of poses.

Of course, two HMMs are equivalent up to a permutation of the (finite) state space. In other words, similar sequences can differ in the order of their poses. We then normalize the ordering of the states by finding for each sequence the state permutation which correspond to the best match between its C matrix and the others.

Even though they have been widely applied to gesture or action recognition, HMMs have been rarely studied as a tool in the gait ID problem [15, 31]. In particular, Kale and Chellappa [17, 19] used the Baum-Welch forward algorithm to compute the log-likelihood of the sequence with respect to a set of learnt Markov models. Given N instances of the gait cycle for the unknown person, a confusion matrix was built from the log-likelihoods, and finally the model having the largest likelihood in the majority of case was chosen.

6. Experiments

We used the CMU Mobo database [12] to test the bilinear approach to view-invariant gaitID. As a matter of fact, as its six cameras are widely separated, it gives us a real chance of testing the effectiveness of the algorithm in a rather realistic setup. In the Mobo database 25 different people perform four different walking-related actions: walking at low speed, walking at high speed, walking along an inclined slope, and walking while carrying a ball. The sequences were acquired indoor, with the people walking on a treadmill at constant speed. The cameras are more or less equally spaced around the treadmill, roughly positioned around the origin of the world coordinate system [12] (see Figure 2). Each sequence is composed by some 340 frames, encompassing 9-10 full walking cycles (left-parallel-right-parallel). We renamed the six cameras originally called 3,5,7,13,16,17 as 1,2,3,4,5,6.

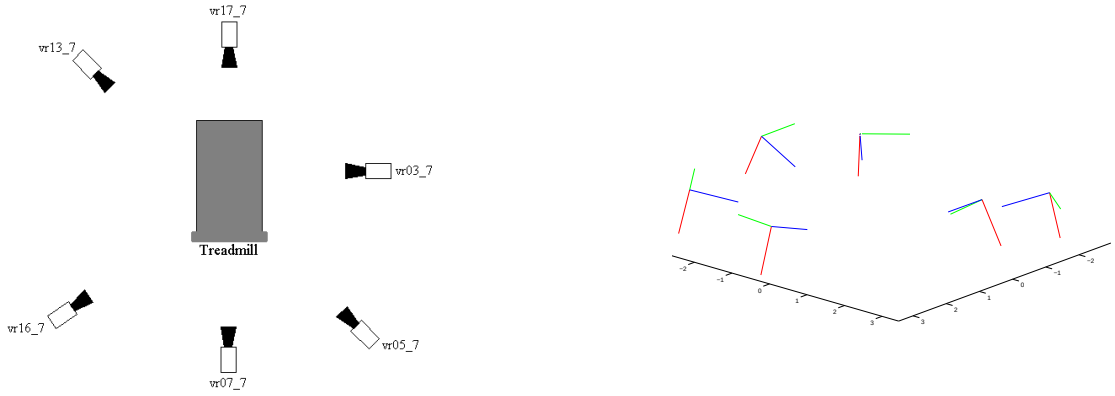

Figure 2: Location and orientation of the six cameras of the CMU Mobo database. The Z,Y, and X axes of the camera reference frames are drawn in blue, green, and red respectively. Scales are in meters.

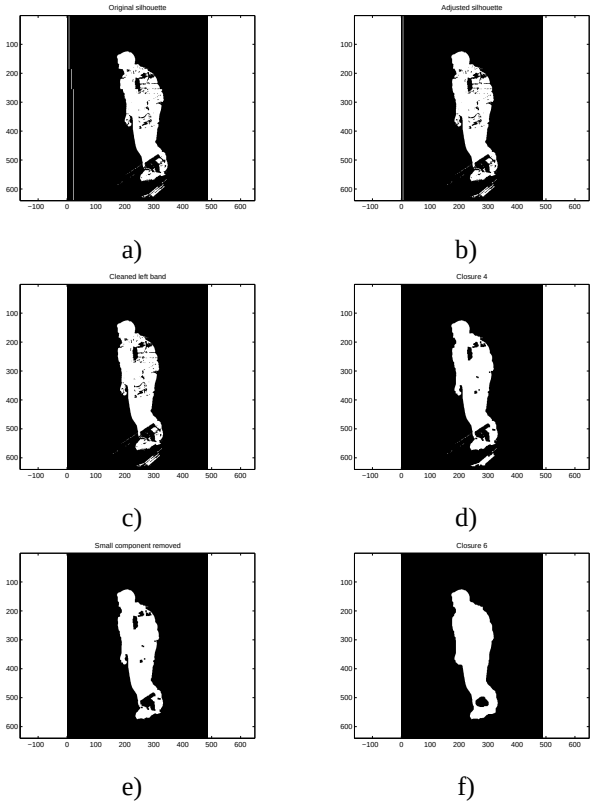

Figure 3: Restoration process of the original silhouettes a). b) First the vertical band on the left is checked to adjust the horizontal shift of each row. Then the left band is removed c), and morphological closure operator is applied a first time d). The smaller connected components are discarded e), and finally the closure operator is applied again to fill the small gaps in the shape f).

6.1. Experiments on ID-view classification

We considered the slow-walk motion only, and set up six different experiments, one for each available camera. In each experiment we used the sequences related to all cameras but one as training set, and built an asymmetric bilinear model

as in Section 3.1. We then used the sequences shot by the remaining camera as test data, and implemented the bilinear classification described in Section 3.2. To get a significantly large dataset, we adopted the period estimation technique in [28] to separate the original long sequences into a larger number of subsequences each spanning three walking cycles. We then computed the feature matrices for each subsequence, and preprocessed them using the HMM-EM algorithm with $n = 3$ states to generate a dataset of pose matrices C, each containing three 10-dimensional pose vectors (2 components for each of the 5 regions of the silhouettes). We finally stacked their columns into a 30-dimensional observation vector for each subsequence. These observation vectors formed our dataset. We used the set of silhouettes provided with the database.

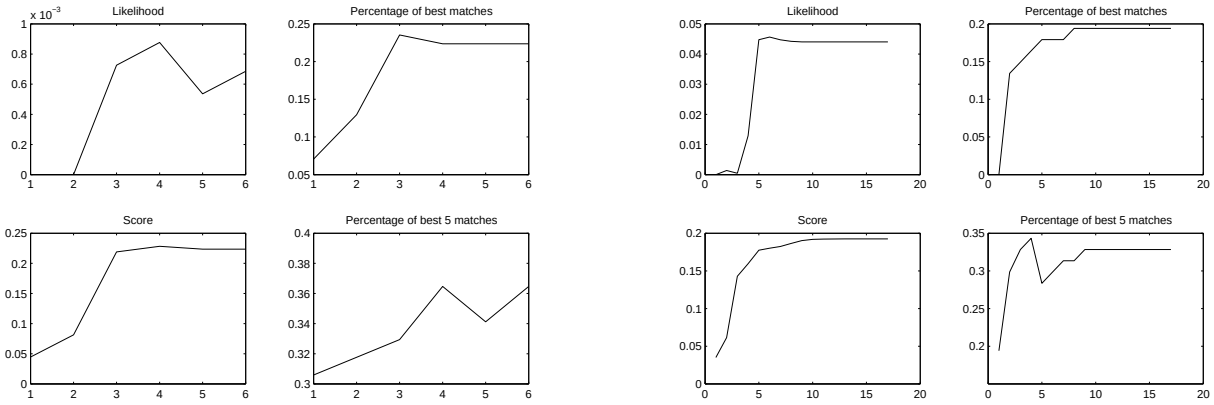

Figure 4: Performance of the EM classification algorithm of Section 3.2 for the experiments with test sequences from camera 4 and 6. The diagrams plot the likelihood and scores along the iterations of the EM algorithm described in Section 3.2. Here the dimension J of the identity space is set to 10.

However, some preprocessing was needed to clean away some artifacts from the original silhouettes, and improve them to increase the chance of getting a good classification. In particular, we adjusted the horizontal shift of each row, and applied morphological operators to close gaps and remove smaller artifacts from the image (Figure 3).

6.2. Performance

The EM algorithm has to be initialized in the E step. We do that by assigning to the probability $p(c|\mathbf{y})$, for each identity (content) label c, a function of the average distance between the test vector $\mathbf{y}$ and all the training vectors extracted from sequences related to the same person c. Namely,

$$p(c|\mathbf{y}) \propto \frac{\max_{\mathbf{v}\in T} \|\mathbf{v}-\mathbf{y}\| - \frac{1}{|T_c|}\sum_{\mathbf{v}\in T_c} \|\mathbf{v}-\mathbf{y}\|}{\max_{\mathbf{v}\in T} \|\mathbf{v}-\mathbf{y}\| - \min_{\mathbf{v}\in T} \|\mathbf{v}-\mathbf{y}\|}$$

where T is the training set, T_c the training vectors of identity c. Figure 4 shows an example of how the EM algorithm for bilinear models performs, in particular in the experiment in which the test sequences come from camera 4. As we know the ground truth identities of all the test sequences, we can compute a "score" of the classification (third plot) as $\sum_{\mathbf{y}\in Y}\sum_{c=1}^{25} p(c|\mathbf{y})\cdot\delta_{\mathbf{y}}^{c}$, where Y is the training set and $\delta_{\mathbf{y}}^{c} = 1$ iff $\mathbf{y}$ has identity c. We can also calculate the percentage of correct top matches (second plot) and the percentage of test sequences for which the correct identity is one of the first 5 matches (fourth plot).
As it is apparent from Figure 4, the behaviors of log likelihood and classification scores are not necessarily correlated: this was originally pointed out in [33]. We then need to learn the optimal number of iterations to get the best score.

6.3. Parameter learning

Actually the bilinear classifier depends on a number of parameters, in particular the variance σ of the mixture distribution (4), and the dimension J of the content space. They can be learned in a second learning stage, by computing the scores of the algorithm for each value of the parameters. Figure 5 illustrates the results of this procedure for the experiment with test

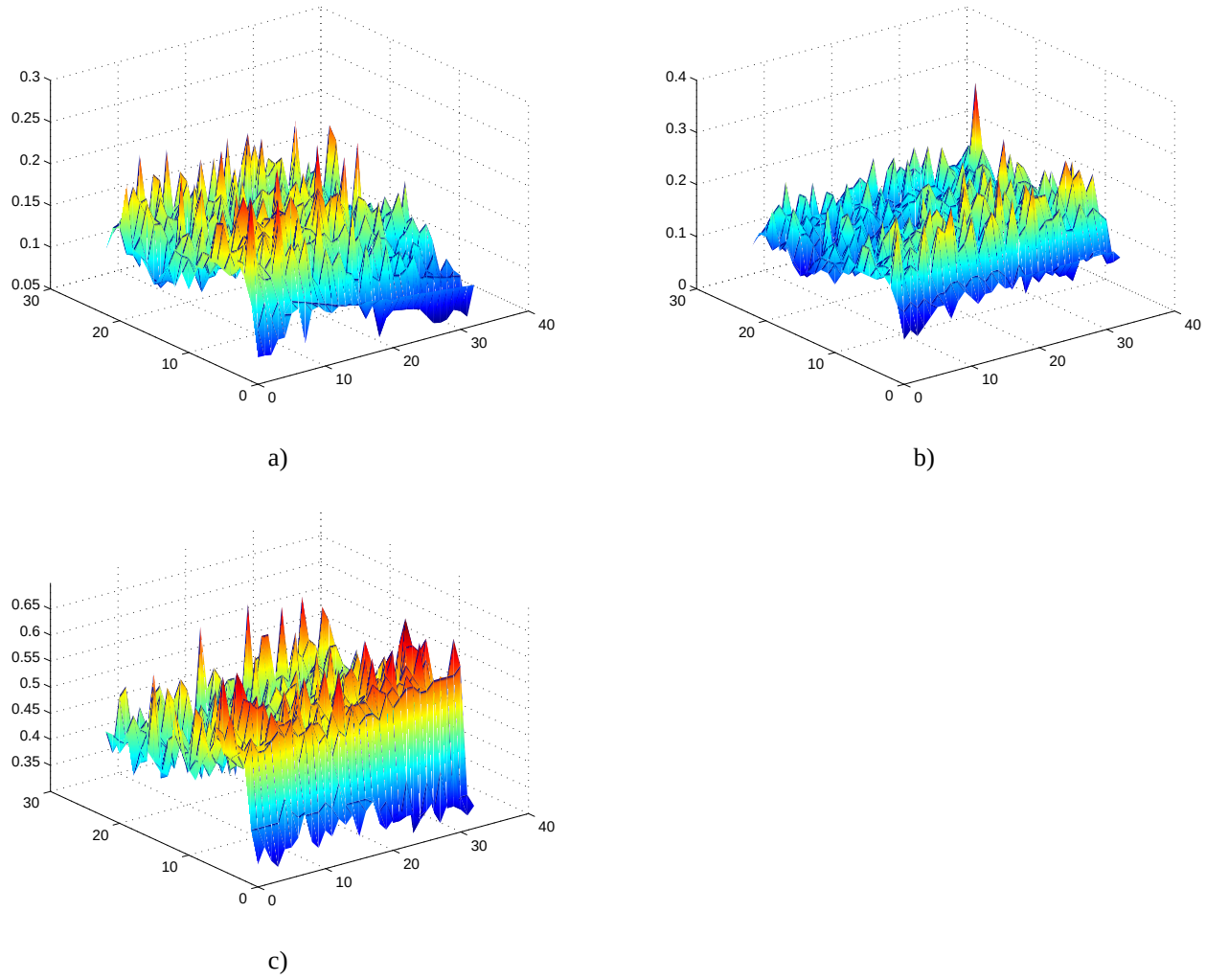

Figure 5: Parameter learning for bilinear models ID-view. The first plot (a) shows the "score" achieved by the classification algorithm for dimension J of the content space varying between 3 and 25 (Y axis), and variance σ of the mixture model varying between 8 and 40 (X axis). The second one (b) shows the percentage of correct top matches on the related test sequences, while the third (c) illustrates the percentage of finding the correct identity in the top 5 matches.

sequences coming from the first view. A visual inspection of Figure 5 suggests how the performance of the algorithm is good for smaller dimensions of the content space, no matter the variance of the mixture model, while it degrades for intermediate values of J. Rather surprisingly, instead, the best top matches are often obtained for high dimensions and high variance $d = 25, \sigma = 40$ (middle). We found this true for all the six experiments. This makes sense as the model need to be allowed a large enough content space to accommodate all the identities. Some differences in the location of the "best" parameters when maximizing the three scores can also be noticed, as $\hat{\sigma}$ and $\hat{J}$ for the percentage of correct top matches or the percentage of correct match in the top 5 do not often coincide.

6.4. Comparison with the baseline algorithm

To get an idea of the comparative performance of our algorithm, we implemented the baseline algorithm described in [28, 27]. The baseline computes similarity scores between a probe sequence S_P and each gallery (training) sequence S_G by pairwise

frame correlation. Namely, a similarity score

$$Sim(S_P, S_G) = median_k(\max_l Corr(S_{P_k}, S_G)(l))$$

is computed for each pair (S_P, S_G), where

$$Corr(S_{P_k}, S_G)(l) = \sum_{j=0}^{Ngait-1} Framesim(S_P(k+j), S_G(l+j))$$

and

$$Framesim(S_P(i), S_G(j)) = \frac{Num(S_P(i)) \cap Num(S_G(j))}{Num(S_P(i)) \cup Num(S_G(j))}$$

is the similarity measure between frames, computed as the ratio between the number of pixel in the intersection of two silhouettes and the number of pixel of their union. $Ngait$ is the estimated length of the cycles, and k is the number of probe subsequences.

Figure 6 compares the results of the bilinear classification with the results of the baseline algorithm for all the six experiments. Both top 1 (top) and top 5 (bottom) matches are compared. It can be easily appreciated how the structure introduced

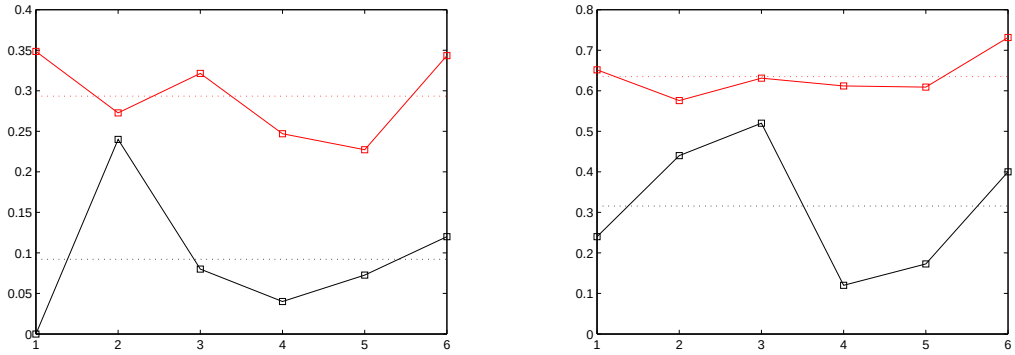

Figure 6: Comparison of the performances of the bilinear approach and the baseline algorithm on the six experiments, one for each view. Top: percentage of correct top matches in the bilinear (red) and baseline (black) algorithms. Bottom: percentage of correct identification in the top 5 matches. The average performances of the two algorithms are drawn in dashed lines.

by the bilinear model into to dataset improves greatly the identification performance, rather homogeneously over all views. The recognition levels are well above chance (4% and 20% respectively). Notice also that the preprocessing stage does not affect the performance of the bilinear classifier. The baseline algorithm instead seems to work better for sequences coming from cameras 2 and 3, which have rather close viewpoints, while it delivers the worst results for camera 1, the most isolated from the others [12]. No particular influence of the silhouette quality could be inferred.

7. Towards bilinear meta-models

In this paper we studied one of the major covariates of the problem of recognizing the identity of humans from gait, i.e. viewpoint dependence. View-invariance is a critical issue to make identity recognition from gait a valid alternative to other biometric systems. We proposed to learn and implement bilinear models in the two major attributes of the input sequences, identity and view, and designed a system in which hidden Markov models with a fixed number of states are used to cluster each sequence into a fixed number of poses to generate the observation data for an asymmetric bilinear model. We used the CMU Mobo database [12] to set up an experimental comparison between the bilinear approach and the standard baseline algorithm, showing how even using simple feature representations of the images they can improve the performance of the recognition from unknown viewpoints.

In a sense, in this paper we applied bilinear models to hidden Markov models with fixed dynamics. In perspective, in situation in which the dynamics is not trivial, for instance when the speed is not constant [32], the application of classification

of dynamical models is an attractive possibility. In particular, this will possibly allow to classify people's identity when executing different arbitrary actions. A challenging issue will then be a choice for the representation of the model parameters. In the near future it will then be worth to extend our analysis to identity recognition from different actions, using bilinear ID-action models.

References

[1] C. B. Abdelkader, R. Cutler, H. Nanda, and L. Davis, "Eigengait: motion-basedrecognition using image self-similarity," *Proceedings of the Audio- and Video-Based Biometric Person Authentication*, Halmstaadt, Sweden, 2001, p. 289-294.

[2] C. BenAbdelkader, R. Cutler, and L.S. Davis, "Stride and cadence as a biometric in automatic person identification and verification," *Proc. AFGR'02*, pp. 357–362, 2002.

[3] B. Bhanu and J. Han, "Individual recognition by kinematic-based gait analysis," *ICPR02*, Vol. 3, pp. 343–346, 2002.

[4] P. Cattin, D. Zlatnik, and R. Borer, "Sensor fusion for a biometric system using gait," *International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 233–238, 2001.

[5] R. T. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait," *IEEE Conf. Automatic Face and Gesture Recognition*, pp. 351-356, 2002.

[6] D. Cunado, J. M. Nash, M. S. Nixon, and J. N. Carter, "Gait extraction and description by evidence-gathering," *Proceedings of AVBPA99*, pp. 43–48, 1999.

[7] D. Cunado, M. S. Nixon, and J. N. Carter, "Automatic extraction and description of human gait models for recognition purposes," *Computer Vision and Image Understanding*, Vol. 90, no. 1, pp. 1–41, April 2003.

[8] N. Cuntoor, A. Kale, and R. Chellappa, "Combining multiple evidences for gait recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003.

[9] 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 B*, Vol. 39, pp. 1–38, 1977.

[10] A. Elgammal and C. S. Lee, "Separating style and content on a nonlinear manifold," *CVPR'04*, June-July 2004.

[11] J. P. Foster, M. S. Nixon, and A. Prgel-Bennett, "Automatic gait recognition using area-based metrics," *Pattern Recogn. Lett.*, Vol. 24, no. 14, pp. 2489–2497, 2003.

[12] R. Gross and J. Shi, "The CMU motion of body (Mobo) database," Tech. report, Carnegie Mellon University, 2001.

[13] J. Han and B. Bhanu, "Statistical feature fusion for gait-based human recognition," *CVPR04*, Vol. 2, pp. 842–847, 2004.

[14] J. B. Hayfron-Acquah, M. S. Nixon, and J. N. Carter, "Automatic gait recognition by symmetry analysis," *Pattern Recogn. Lett.*, Vol. 24, no. 13, pp. 2175–2183, 2003.

[15] Q. He and C. Debrunner, "Individual recognition from periodic activity using hidden Markov models," *IEEE Workshop on Human Motion*, 2000.

[16] A. Y. Johnson and A. F. Bobick, "A multi-view methodfor gait recognition using static body parameters," *Third Proceedings of the Audio- and Video-Based Biometric Person Authentication*, pp. 301-311, Halmstaadt, Sweden, 2001.

[17] A. Kale, A. N. Rajagopalan, N. Cuntoor, and V. Kruger, "Gait-based recognition of humans using continuous HMMs," *IEEE Conf. on Automatic Face and Gesture Recognition*, pp. 321-326, 2002.

[18] A. Kale, A. K. Roy-Chowdhury, and R. Chellappa, "Towards a view invariant gait recognition algorithm," *AVSBS03*, pp. 143–150, 2003.

[19] A. Kale, A. Sunsaresan, A. N. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa, "Identification of humans using gait," *IEEE Trans. PAMI*, Vol. 13, no. 9, pp. 1163–1173, 2004.

[20] C.-S. Lee and A. Elgammal, "Gait style and gait content: bilinear models for gait recognition using gait re-sampling," *AFGR'04*, pp. 147–152, 17-19 May 2004.

[21] L. Lee and W. Grimson, "Gait analysis for recognition and classification," *International Conference on Automatic Face and Gesture Recognition*, pp. 155-162, 2002.

[22] J. Little and J. Boyd, "Recognising people by their gait: the shape of motion," *IJCV*, Vol. 14, no. 6, pp. 83-105, 1998.

[23] Y. Liu, R. Collins, and Y. Tsin, "Gait sequence analysis using frieze patterns," *European Conference on Computer Vision*, Vol. 2, pp. 657-671, 2002.

[24] R. Elliot, L. Aggoun and J. Moore, *Hidden Markov models: estimation and control*, 1995.

[25] H. Murase and R. Sakai, "Moving object recognition in eigenspace representation: gait analysis andlip reading," *Pattern Recognition Lett.*, Vol. 17, pp. 155-162, 1996.

[26] S. A. Niyogi and E. H. Adelson, "Analyzing and recognizing walking figures in XYT," *IEEE Proceedings of Computer Vision and Pattern Recognition*, pp. 469-474, 1994.

[27] P. J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. W. Bowyer, "The gait identification challenge problem: data sets and baseline algorithm," *International Conference on Pattern Recognition*, Vol. 1, pp. 385–388, 2002.

[28] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. and Mach. Intel.*, Vol. 27, pp. 162–177, Feb. 2005.

[29] G. Shakhnarovich, L. Lee, and T. Darrell, "Integrated face and gait recognition from multiple views," *Proceedings of CVPR'01*, pp. 439–446, 2001.

[30] J. Shutler, M. Nixon, and C. Harris, "Statistical gait recognition via velocity moments," *Proc. IEE Colloquium on Visual Biometrics*, pp. 10/110/5, March 2000.

[31] A. Sundaresan, A. K. Roy-Chowdhury, and R. Chellappa, "A hidden Markov model based framework for recognition of humans from gait sequences," *ICIP'03*, Vol. 2, pp. 93–96, 2003.

[32] R. Tanawongsuwan and A. Bobick, "Modelling the effects of walking speed on appearance-based gait recognition," *CVPR04*, Vol. 2, pp. 783–790, 2004.

[33] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, Vol. 12, pp. 1247-1283, 2000.

[34] D. Tolliver and R. Collins, "Gait shape estimation for identification," *International Conference on Audio- and Video-Based Biometric Person Authentication*, pp.734-742, 2003.

[35] R. Urtasun, and P. Fua, "3D tracking for gait characterization and recognition," Technical Report IC/2004/04, Computer Vision Laboratory, Swiss Federal Institute of Technology, 2004.

[36] I. Robledo Vega and S. Sarkar, "Representation of the evolution of feature relationship statistics: Human gait-based recognition," *IEEE Trans. PAMI*, Vol. 25, pp. 1323-1328, 2003.

[37] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. Pattern Anal. and Mach. Intel.*, Vol. 25, pp. 1505-1518, 2003.

[38] C. Y. Yam, M. S. Nixon, and J. N. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognition*, Vol. 37, no.5, pp. 1057–1072, 2004.