Using Bilinear Models for View-invariant Action and Identity Recognition

Fabio Cuzzolin
UCLA Vision Lab, University of California at Los Angeles
Boelter Hall, 90025 Los Angeles, CA
cuzzolin@cs.ucla.edu

Abstract

Human identification from gait is a challenging task in realistic surveillance scenarios in which people walking along arbitrary directions are imaged by a single camera. In this paper, motivated by the view-invariance issue in the human ID from gait problem, we address the general problem of classifying the "content" of human motions of unknown "style". Given a dataset of sequences in which different people walking normally are seen from several well-separated views, we propose a three-layer scheme based on bilinear models, in which image sequences are mapped to observation vectors of fixed dimension using Markov modeling. We test our approach on the CMU Mobo database, showing how bilinear separation outperforms other approaches, opening the way to view- and action-invariant identity recognition, as well as subject- and view-invariant action recognition.

1. Introduction

Biometrics has received 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 be identified 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 [33], starting from a seminal work of Niyogi and Adelson [25].

A variety of techniques have been introduced, mostly based on silhouette analysis [4, 34]. Many other gait signatures, however, have been studied, ranging from optical flow [22], to velocity moments [28] shape symmetry [14], frieze patterns [23], height and stride estimation, static body parameters [16], area-based metrics [10], or multiple features [3, 27, 6]. Concerning classification, a number of methods apply some pattern recognition technique after dimensionality reduction (through eigenspaces [1, 24], or PCA/MDA [31, 13]). Others employ stochastic models like hidden Markov models (HMMs) to describe gait dynamics [15].

In the last few years, a number of public databases have been made available which can be used as a common ground on which validate the algorithms. The USF database [26], for instance, has been designed to study the effect of many factors (covariates) on identity classification in a realistic, outdoor context with cameras located at a distance. However, the experiments contemplate only two cameras at fairly close viewpoints (30 degrees or so apart), while people are imaged while walking along the opposite side of an ellipse, so that the resulting views are almost fronto-parallel. Appearance-based algorithms work well in the experiments concerning viewpoint variability, but otherwise for widely separated views. In a realistic setup, the person to identify would walk in the surveyed area from an arbitrary direction. View-invariance [32, 35, 2, 17, 27, 16] is then a crucial issue to make identification from gait suitable for real-world applications.

1.1. View-invariance in gaitID

This problem has been actually studied in the gait ID context by many people. If a 3D articulated model of the moving person is available, tracking can be used as a pre-processing stage to drive recognition. Cunado et al. [5], for instance, used their evidence gathering technique to analyze the leg motion in both walking and running gaits, providing estimates of the inclination of thigh and leg. Yam et al. [35] also worked on a similar model-based approach. Urtasun and Fua [32] proposed an approach to gait analysis that relies on fitting 3D temporal motion models to synchronized video sequences, while Bhanu and Han [2] matched a 3D kinematic model to 2D silhouettes.

Model-based 3D tracking, however, is a difficult task, as manual initialization is often required, and optimization in a high-dimensional parameter space is sensitive to convergence defects. Kale et al. [17] proposed instead a method to generate a synthetic side-view of the moving person us-

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

ing a single camera. Shakhnarovich et al. [27] suggested a view-normalization technique in a multiple camera context, using the volumetric intersection of the visual hulls. Johnson e Bobick also presented a multi-view gait recognition method using static body parameters recovered across multiple views [16].

1.2. From view-invariance to style-invariance

View-invariance can be seen as a particular case of more general issue. Consider a dataset of observations possessing more than a single categorical label: For instance a database of human movements. Each motion can in fact be classified according to the person who performed it, the category of action performed (i.e. walking, reaching out, pointing, etc.), or (if the number of cameras is finite) the viewpoint from which the sequence is shot.

This situation is naturally described in terms of multi-linear models. Bilinear models, in particular [30], can be seen as a tool for separating "style" and "content" of the objects to classify, meaning two distinct class labels of the same objects. As they are capable to learn how factors interact in such a mixed training set, bilinear models allow for instance 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. Analogously, other important vision problems like identity recognition from *unknown actions*, or again *view-invariant* or *identity-invariant* action recognition can be addressed in terms of bilinear classification.

Therefore we here propose a *three-layer model* in which each motion sequence is considered as an observation depending on three factors (*identity*, *action* type, and *view*) from which a bilinear model can be trained by considering two of those factors at a time. While in the first layer features are extracted from a single image, in the second level each feature sequence is given as input to a Markov model. Assuming fixed dynamics, the HMM would cluster the movement into a fixed number of poses. The stacked vector of these poses would then form an observation vector representing the sequence. After learning a bilinear model for such a set of observations, we can then classify (determine the content of) new sequences characterized by a different style label. In the final section we will show experiments on the CMU Mobo database concerning ID and action recognition, showing how this approach performs significantly better than other known approaches.

2. Bilinear models

Bilinear models have been introduced by Tenenbaum et al. [30] 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_{ijk}^s 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 $\{a_{jk}^s = \sum_i w_{ijk}^s a_i^s\}$, a style-specific linear map from the content space to the observation space.

2.1. 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, $\mathbf{A} = [\mathbf{A}^1 \cdots \mathbf{A}^S]'$, $\mathbf{B} = [\mathbf{b}^1 \cdots \mathbf{b}^C]$. 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 [30].

2.2. 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 solved by fitting a mixture model to the learned bilinear model by means of the EM algorithm [7]. 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

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

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

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

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 critically depends on the goodness of the representation chosen for the observation vectors. However, it was originally presented as a way of finding *approximate* solutions to problems in which two factors are involved [30], without precise context-based knowledge. In the gaitID context, for instance, Elgammal and Lee have analyzed the geometry of cycles in the visual space, and adopted local linear embedding [30] as a tool to re-sample homogeneously each cycle into a fixed number of poses.

[1]The general formulation allows the presence of more than one unknown style, [30].

3. A three-layer model

As we mentioned above, in human motion analysis movements can be characterized by a number of different labels: each motion can in fact be classified according to the identity of the person, the category of action performed (i.e. walking, reaching out, pointing, etc.), or (if the number of cameras is finite) the viewpoint from which the sequence is shot.
They hence naturally fall in a context of multilinear modeling, in which a dataset of observations can be thought of as a linear mixture driven by two or more factors.
Elgammal and Lee have recently used them to separate pose and ID from a database of poses in the context of GaitID [20, 8]. Here we propose the use of bilinear models to represent and classify movements regardless the "style" with which they are executed. In more practical terms, this allows us to address problems like *view-invariant identity recognition*, identity recognition from *unknown* gaits, classification of actions from *unknown viewpoints* or performed by *new persons*.

We designed a *three-layer model* in which each motion sequence is considered as an observation depending on three factors (*identity*, *action* type, and *view*) from which a bilinear model can be trained by considering two of those factors at a time. We can then apply the technique of Section 2.2 to classify motions regardless their style, as we will see in Section 4.

3.1. First layer: feature representation

We chose a simple but effective feature representation of the silhouettes to reduce the computational load. In particular, given a silhouette we detect its center of mass, rescale it to the associated bounding box, and project its contours on one or more lines passing through the center of mass (see Figure 1). We chose this after testing a number of different representations: among those the principal axes of the body-parts as they appear in the image [21], size functions [11], and PCA representation of the contours. Especially the latter turned out to be rather unstable.

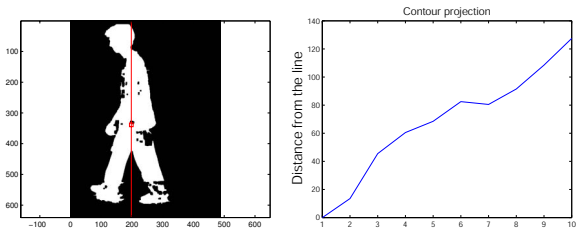

Figure 1. Feature extraction. Left: a number of lines is drawn through the center of mass of the silhouette. Right: the distance of the points on the contour from the line is computed: these values for all the lines form the feature vector.

3.2. Second layer: HMMs as sequence descriptors

If the contour is projected onto 2 orthogonal lines, and we set 10 components for each projection, each image is then represented by a 40-dimensional feature vector. Image sequences are then encoded as sequences of feature vectors, in general of different duration. To make them suitable inputs for a bilinear model learning stage (Section 2.1) we need to find a homogeneous representation. Hidden Markov models [9] provide us with a tool for transforming each sequence into a fixed-length observation vector [2] .

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 [9] 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 [3] 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 (see [9] for the details).

3.2.1 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 A 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 poses are automatically associated with 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].

As in the gait ID case the dynamics is the same for all the sequences (as all of them are instances of the walking motion) it can be factored out: The topology of the resulting HMM is constant. If we also assume that people are walking at constant speed, the transition matrix A is of no use and can be neglected. Hence the matrix C of the poses can be used as a descriptor for the entire sequence [4].

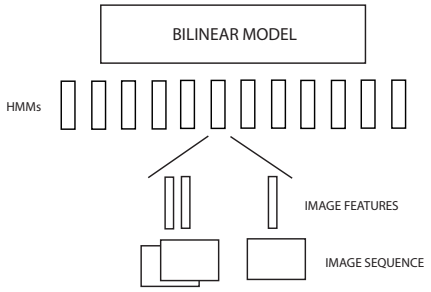

Figure 2. Scheme of the three-layer model for collections of multiple-label movements. Features (bottom layer) are extracted from each image of the sequence. Those feature vectors are fed to an HMM with a fixed number of states, yielding a dataset of Markov models, one for each sequence (second layer). The stacked versions of the C matrices of these models are then used as input vectors for the bilinear model (top layer).

3.3. Third layer: bilinear model of HMMs

The pose matrix C of each sequence can eventually be stacked into a single observation vector representing the sequence. The dataset of sequences is then encoded as a dataset of these observation vector of homogeneous length, even though the original sequences had different duration. They can then be used to build a bilinear model for a dataset of human motions. The procedure can then be summarized as follows:

- each image sequence is transformed into a sequence of feature vectors;
- those feature sequences are fed to EM algorithm, yielding an N-state HMM for each movement;
- the (pose) C matrix of each HMM is stacked to yield on observation vector;
- the algorithm of Section 2.1 is used to build an asymmetric bilinear model for the dataset.

The three-layer model is depicted in Figure 2. Given a dataset of motions, we can use this algorithm to built an asymmetric bilinear model from the sequences related to all style labels but one (training set). We can then use a bilinear classifier (Section 2.2) to label the the sequences associated with the remaining style (testing set), as we will see in the rest of the paper.

[2] 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, 29]. In particular, Kale and Chellappa [18] used the Baum-Welch forward algorithm to compute the log-likelihood of the sequence with respect to a set of learnt Markov models.

[3] Note that A, C here have nothing in common with $\mathbf{A}, \mathbf{C}$ of Section 2.

[4] 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.

4. Experiments

We used the CMU Mobo database [12] to extensively test the bilinear approach to gaitID and action recognition. As its six cameras are widely separated, it gives us a real chance of testing 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]. 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.

4.1. From view-invariant gaitID to ID-invariant action recognition

The sequences of the Mobo database have then three different labels: identity, action, and viewpoint. We then set up four series of test in which we built bilinear models by selecting a content label and a style label among the three, respectively: content=ID, style=view (*view-invariant gaitID*); content=ID, style=action (*action-invariant gaitID*); content=action, style=view (*view-invariant action recognition*); content=action, style=ID (*style-invariant action classification*). The remaining factor can then be considered as a nuisance. In each experiment we used the sequences related to all the style labels but one as training set, and built an asymmetric bilinear model as in Section 2.1. We then used the sequences associated with the remaining style as test data, and implemented the bilinear classifier (Section 2.2).

To get a significantly large dataset, we adopted the period estimation technique in [26] to separate the original long sequences into a larger number of subsequences each spanning three walking cycles. This way we obtained a collection of 2080 sequences, roughly equally divided among the six views, the 25 IDs, and the four actions. We then computed feature matrices for each subsequence, and applied the HMM-EM algorithm with $n = 2$ states to generate a dataset of pose matrices C, each containing two pose vectors as columns. We finally stacked those columns into a single observation vector for each subsequence. These observation vectors would finally form our dataset. We used the set of silhouettes provided with the database, after some preprocessing to clean away artifacts from the original images. In the following we will measure the performance of the algorithm using both the percentage of correct best matches and the percentage of test sequences for which the correct identity is one of the first $k = 3$ matches.

The bilinear classifier depends on a small number of parameters (Section 2.2), in particular the variance σ of the mixture distribution (4), and the dimension J of the content space. They can be learned in a learning stage, by computing the score of the algorithm when applied to the training set for each value of the parameters. Basically the model needs to be allowed a large enough content space to accommodate all the content labels. Most important of all, however, is the initial value of the probability $p(c|\mathbf{y})$ with which each test vector $\mathbf{y}$ belongs to a content class c. Again, this can be learned from the training set by maximizing the classification performance using some sort of *simulated annealing* technique to overcome local maxima.

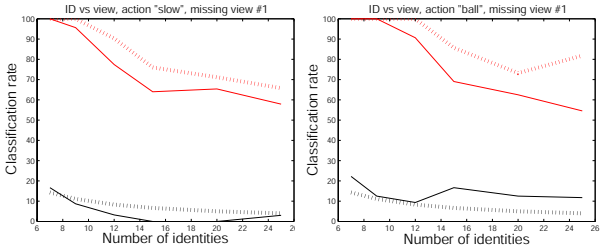

Figure 3. View-invariant gaitID for sequences related to the same action: "slow" (left) and "ball" (right). View 1 is used as test view, and the classifcation rate is plotted for increasing numbers of identities in the dataset. The percentage of correct best matches is shown in solid red, while the rate of a correct match in the first 3 is plotted in dotted red. For comparison the performance of the KL-nearest neighbor classifier on the dataset of HMMs is shown in solid black (pure chance is also plot in dashed black).

4.2. Identity versus view

In the first series of tests we considered "identity" as content label and "viewpoint" as style label. This way we could test the view-invariance of a gaitID bilinear classifier. We report here the results of different kinds of tests. In Figure 3 we selected the subset of the dataset associated with a single action (nuisance). We then measured the performance of the bilinear classifier for view-invariant gaitID using view 1 as test view, for an increasing number of persons in the dataset. To get an idea of the comparative performance of our algorithm, we implemented a simple nearest neighbor classifier which assigns to each test sequence the identity of the closest Markov model (using the standard Kullback-Leibler divergence [19]). Figure 3 clearly shows how the bilinear classifier greatly outperforms a naive NN classification of the Markov models built from the sequences. The depressing results of the KL-NN approach testifies the difficulty of the task.

Figure 4 instead compares the two algorithms as the test viewpoint varies, for the two sub-datasets related to the actions "slow" and "ball" with 12 identities. Again the NN-

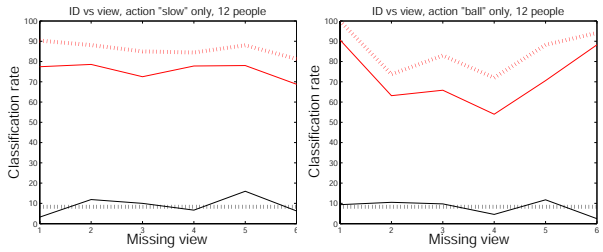

Figure 4. View-invariant gaitID for sequences related to actions "slow" (left) and "ball" (right), and different selection of the test view (from 1 to 6). The first 12 identities are considered. Colors as above.

KL classifier performs around pure-chance levels, while the bilinear classifier reaches excellent scores of some 90% for some views. Relatively large deviations in the second plot are due, according to our experience, to the parameter learning algorithm being stuck to a local maximum. Figure 5-left illustrates the performance of the algorithm as a function of the nuisance factor, i.e. the performed action: ball=1, fast=2, incline=3, slow=4. The classifier does not exhibit any particular dependence. We also implemented

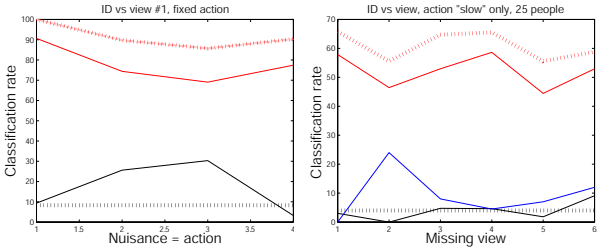

Figure 5. Performance of the bilinear classifier in the ID vs view experiment. Left: score as a function of the nuisance (action), test view 1. Right: score for the dataset of sequences related to the action "slow", and different selection of the test view (from 1 to 6), 25 identities. The classification rate of the baseline algorithm is in blue: the other colors are as above.

for sake of comparison the *baseline algorithm* described in [26], which basically computes similarity scores between a probe sequence S_P and each gallery (training) sequence S_G by pairwise frame correlation. Figure 5-right compares the results of the bilinear classification with the results of both the baseline algorithm and the KL-based approach for all the six possible test views, in the complete dataset with 25 identities. It can be easily appreciated that the structure introduced by the bilinear model improves greatly the identification performance, rather homogeneously over all the views. 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]. The KL-based nearest neighbor approach is not distinguishable from pure chance.

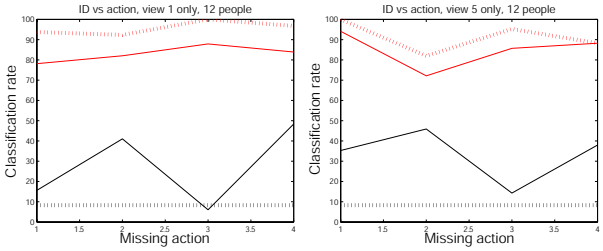

Figure 6. Action-invariant gaitID for sequences related to viewpoints 1 (left) and 5 (right), and different selection of the test action (from 1 to 4). The first 12 identities are considered. Colors as in Figure 4.

4.3. Identity versus action

In a different experiment we validated the assumption that a person can be recognized *even from an action he never performed before*, provided that we have seen this action performed by other people. In our case the assumption is quite reasonable, since all the actions in the database are nothing but different variations on the gait gesture. We then built bilinear models for content=ID, style=action from a training set of sequences related to three actions, and classified the remaining sequences (instances of the fourth action) using the bilinear method. Figures 6 to 8 support the ability of bilinear classification to allow identity recognition even from different gestures.

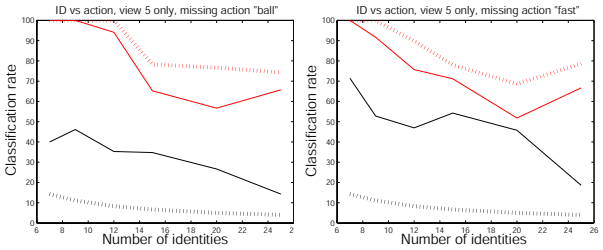

Figure 7. Action-invariant gaitID. In the left diagram sequences related to viewpoint (nuisance) #5 are considered, and "ball" is used as missing action (test style). In the right diagram sequences related to viewpoint #4 are considered, and "fast" is used as test (missing) action. The classifcation rate is plotted for increasing numbers of identities, colors as above.

The best-match ratio is around 90% for twelve persons, even though it slightly declines (Figure 7) for larger datasets (the parameter learning algorithm is stopped after a fixed period of time, yielding suboptimal models). The NN-KL classifier performs relatively better in this experiment, but well below an acceptable level. Again, Figure 8 illustrates the various recognition rates as functions of the nuisance (viewpoint).

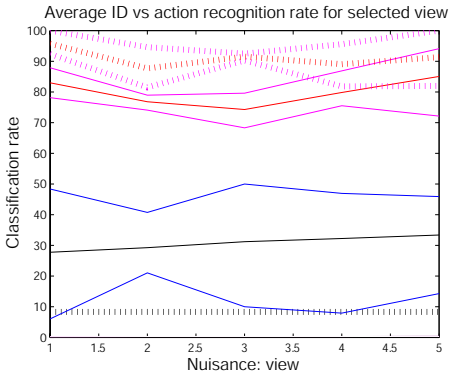

Figure 8. Performance of the bilinear classifier in the ID vs action experiment as a function of the nuisance (view=1:5), averaged over all the possible choices of the test action. The average best-match performance of the bilinear classifier is shown in solid red, with minimum and maximum scores in solid magenta. The related best-3 matches ratio is plotted in dotted red, minimum and maximum scores in dotted magenta. The average performance of the KL-nearest neighbor classifier is shown in solid black, minimum and maximum in blue. Pure chance is in dashed black.

4.4. Action versus identity

In the third experiment we considered the situation in which we know a database of action performed by a number of persons, and we want to recognize one of those known actions when performed by an *unknown* person (style-invariant action classification). In this case content = action, style = identity, while the viewpoint from which the image sequences are shot is a nuisance factor. Figure 9-left shows the performance of the bilinear method on the *entire* dataset, regardless the viewpoint, for all the possible choices of the test identity. The scores obtained with two different image feature representations are show: contour projection as in Section 3.1, and principal axes of fixed regions of the silhouette [27, 16, 21]. [5] Again, Figure 9-right shows the results as function of the nuisance (viewpoint).

4.5. Action versus view

In a final experiment we considered the problem of recognizing a *known action* performed by *a view not in the dataset*. In this case the content label is "action", the style label is "view", and identity is the nuisance factor. Figure 10-left shows the correct classification rate on the entire dataset, regardless the identity of the walking person, for all the possible choices of the test viewpoint. The scores obtained with the two mentioned features are shown, proving how feature choice is essential for the performance of the three-layer model. A comparison with the results obtained

[5] Since we are working with only four actions (even though very similar to each other) it makes little sense to report the percentage of best $k = 3$ matches.

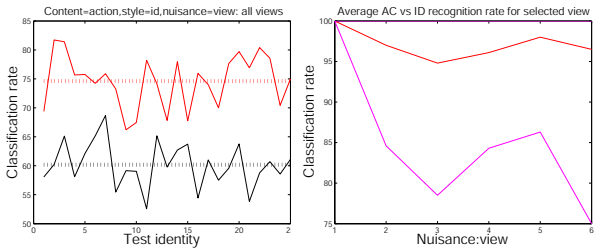

Figure 9. Identity-invariant action recognition using the bilinear classifier. Left: the entire dataset is considered, regardless the nuisance label (the viewpoint from which the sequence is shot). The correct classification percentage is shown as a function of the test identity in black (for models using Lee's features) and red (contour projections). Related mean levels are drawn as dotted lines. Right: recognition rate as a function of the nuisance (view=1:6), averaged over all the possible choices of the test identity. The average best-match performance of the bilinear classifier is shown in red, with minimum and maximum scores in magenta.

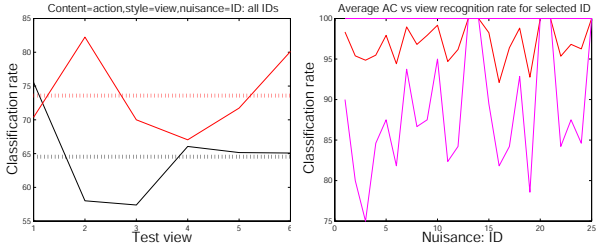

Figure 10. View-invariant action recognition. Left: the entire dataset is considered, regardless the nuisance label (the identity of the person performing the action). The correct classification percentage is shown as a function of the missing view in black (for models using Lee's features) and red (contour projections). Related mean levels are drawn as dotted lines. Right: recognition rate as a function of the nuisance (ID=1:25), averaged over all the possible choices of a test view. The average best-match performance of the bilinear classifier is shown in red, with minimum and maximum scores in magenta.

for fixed-nuisance datasets (Figure 10-right) confirms that the variability induced by a third factor disturbs the performance of the algorithm.

To conclude, it is worth to mention that the bilinear classifier itself is almost instantaneous. However, the parameter learning algorithm can typically take a few minutes as simulated annealing works its way towards the optimal solution.

5. Conclusions

In this paper, motivated by the view-invariance issue in the gaitID problem we addressed the problem of classifying multiple-label motions. We designed a three-layer model 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 other standard algorithms in several experiments ranging from view-invariant gaitID to identity-invariant action recognition, showing how bilinear modelling can improve recognition performances when the test motion is performed in an unknown style.

Acknowledgements

Research supported by ONR N00014-03-1-0850:P0001 and AFOSR E-16-V91-G2

References

[1] C. B. Abdelkader, R. Cutler, H. Nanda, and L. Davis. Eigengait: motion-basedrecognition using image self-similarity. *Proc. of AVBPA'01, Halmstaadt, Sweden*.

[2] B. Bhanu and J. Han. Individual recognition by kinematic-based gait analysis. In *Proc. ICPR02*, volume 3, pages 343–346, 2002.

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

[4] R. T. Collins, R. Gross, and J. Shi. Silhouette-based human identification from body shape and gait. In *AFGR'02*, pages 351–356, 2002.

[5] D. Cunado, J. M. Nash, M. S. Nixon, and J. N. Carter. Gait extraction and description by evidence-gathering. In *Proc. of AVBPA99*, pages 43–48, 1999.

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

[7] 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*, 39.

[8] A. Elgammal and C. S. Lee. Separating style and content on a nonlinear manifold. In *CVPR'04*, 2004.

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

[10] J. P. Foster, M. S. Nixon, and A. Prgel-Bennett. Automatic gait recognition using area-based metrics. *Pattern Recogn. Lett.*, 24(14):2489–2497, 2003.

[11] P. Frosini. Measuring shape by size functions. In *Proceedings of SPIE on Intelligent Robotic Systems*, volume 1607, pages 122–133, 1991.

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

[13] J. Han and B. Bhanu. Statistical feature fusion for gait-based human recognition. In *CVPR'04*, volume 2, pages 842–847, 2004.

[14] J. B. Hayfron-Acquah, M. S. Nixon, and J. N. Carter. Automatic gait recognition by symmetry analysis. *Pattern Recogn. Lett.*, 24(13):2175–2183, 2003.

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

[16] A. Y. Johnson and A. F. Bobick. A multi-view methodfor gait recognition using static body parameters. In *Proc. of AVBPA'01*, pages 301–311, 2001.

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

[18] 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*, 13(9):1163–1173, 2004.

[19] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Math. Stat.*, 22:79–86, 1951.

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

[21] L. Lee and W. Grimson. Gait analysis for recognition and classification. In *AFGR'02*, pages 155–162, 2002.

[22] J. Little and J. Boyd. Recognising people by their gait: the shape of motion. *IJCV*, 14(6):83–105, 1998.

[23] Y. Liu, R. Collins, and Y. Tsin. Gait sequence analysis using frieze patterns. In *ECCV'02*, volume 2, pages 657–671, 2002.

[24] H. Murase and R. Sakai. Moving object recognition in eigenspace representation: gait analysis andlip reading. *Pattern Recognition Lett.*, 17.

[25] S. Niyogi and E. Adelson. Analyzing and recognizing walking figures in XYT. In *CVPR'94*, pages 469–474, 1994.

[26] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer. The humanID gait challenge problem: Datasets, performance, and analysis. *PAMI*, 27.

[27] G. Shakhnarovich, L. Lee, and T. Darrell. Integrated face and gait recognition from multiple views. In *CVPR'01*, pages 439–446, 2001.

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

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

[30] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12.

[31] D. Tolliver and R. Collins. Gait shape estimation for identification. In *Proc. of AVBPA'03*, pages 734–742, 2003.

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

[33] I. R. Vega and S. Sarkar. Representation of the evolution of feature relationship statistics: Human gait-based recognition. *IEEE Trans. PAMI*, 25.

[34] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *PAMI*, 25.

[35] C. Yam, M. Nixon, and J. Carter. Automated person recognition by walking and running via model-based approaches. *Pattern Recognition*, 37(5):1057–1072, 2004.