Gait Recognition using Dynamic Affine Invariants

Alessandro Bissacco Payam Saisan

UCLA CSD TR040014

Abstract

We present a method for recognizing classes of human gaits from video sequences. We propose a novel image-based representation of human gaits. At any instant of time a gait is represented by a vector of affine invariant moments. These invariants are computed on the binary silhouettes corresponding to the moving body. We represent the time trajectories of the affine invariant moment vector as the output of a linear dynamical system driven by white noise. The problem of gait classification then boils down to formulating distances and performing recognition in the space of linear dynamical systems. Experimental results demonstrate the discriminative power of the proposed approach.

1 Introduction

We live in a dynamic world, constantly analyzing and parsing time varying streams of sensory information. Almost all biological creatures equipped with the sense of vision use dynamic cues to analyze their surrounding for critical survival decisions. Clearly there is an abundance of information embedded in the dynamics of visual signals[1]. In this work we focus on extracting and exploiting the temporal structure of video sequences for the purpose of recognizing human gaits.

By observing a person walking from a distance, we can often tell whether the subject is a human, identify their gender, or make predictions about individual traits like age or physical health. We postulate that such information is encoded not necessarily in the static appearance, but mostly in the *dynamics* of the moving body. In Johansson's experiments [34] one cannot tell much from a single frame, however when the sequence is animated suddenly the scene is easily parsed. Johansson's experiments show that even in the lack of all comprehensible static content, the dynamics of a few moving dots can contain sufficient information to correctly decipher the underlying physical phenomenon.

In this paper we address the problem of recognizing a person walking from one jumping, running, hopping or dancing, and we want to do this independent of the person and her pose. We propose a novel representation of human gaits based on computing affine invariant moments on the binary silhouette of the moving body. Following our previous work [4], we model the dynamic of these affine moments as the output of a linear dynamical system, and define a distance between models for the purpose of recognition.

Our representation is insensitive to a wide range of variabilities in the images, such viewing condition and identity of the performer. Another advantage of our approach is that it does not use any model of the appearance of a person, so it can be naturally extended to other classes of periodic motion. Finally, this method does not require to perform a tracking step, which is usually a challenging task for most gait sequences.

1.1. Previous Work

The problem of image-based human motion analysis and recognition has been receiving considerable attention in the literature. Most of the proposed approaches involve tracking the pose of the human body, represented either as kinematic chain of body parts [11, 16, 21], or as spatial arrangement of blobs [9] or point features [1]. Statistical models, such as standard [5, 1] and parametric [6, 17] Hidden Markov Models are then fitted to the tracking data and likelihood tests are used for recognition. In [10, 12, 2] mixed-state statistical models for the representation of motion have been proposed, and in [18, 19] particle filters have been applied in this framework for estimation and recognition. In [4] linear Gaussian models have been used, and recognition is performed by defining a metric on the space of models.

Other techniques do not require an explicit model of the human body. Zelnik-Manor et al. [15] propose a statistics of the spatio-temporal gradient at multiple temporal scales and use it to define a distance between video sequences.

[1] Naturally there is also a great deal of information in the photometry and geometry of the scene that can be conveyed in a single static frame. However, in this study we concentrate on the scene dynamics.

Some approaches [14, 20, 8] are specific to recognition of periodic motion, such as the human gaits we consider in this paper. In [14] classification is based on periodicities of a similarity measure computed on tracked moving parts. Little and Boyd [8] use Fourier analysis to compute the relative phase of a set of features derived from moments of optical flow, and employ the resulting phase vector for classification. Bobick and Davis [7] propose a description based on the spatial distribution of motion, the Motion Energy and Motion Histogram Images. Recognition is done by comparing Hu moments [22] of those images with a set of stored models. In [13], the problem is recognizing actions from video taken from a distance, where the person appears only as a small patch. They compute a set of spatio-temporal motion descriptors on a stabilized figure-centric sequence, and match the descriptors to a database of preclassified actions using nearest neighbor classification.

2 Extracting an Affine Invariant Representation

An instance of a gait here is an image sequence of about two to three seconds (50-70 frames) long. We assume that the sequence contains a human subject performing an action like walking or running. In their raw pixel form image sequences are far too cluttered with irrelevant information. We are only interested in the part related to the person in the image and more specifically his/her motion. A successful isolation of the dynamics information is highly dependent on the extraction of a representation that is insensitive to such nuisance factors like the background, clothing, lighting and viewing angle. While well established techniques with arbitrary degrees of sophistication can be deployed for extracting appearance free representations we note that simple silhouette's are conveniently insensitive to appearance factors[1]. They can be easily extracted from motion sequences using background subtraction techniques. However, we need to be able to recognize a walk not just invariant to appearance, but also to the vantage point. Much work has been done with features like textures, edges, transform coefficients (Fourier, wavelets) and matrix factorizations. To account for perspective distortions and variations of the vantage point we must go beyond these and consider more general statistical features. While an appearance free feature was straightforward to attain, extracting a geometric invariant feature requires some attention. For this we follow well established results from theory of geometric invariance and look at affine invariance. Utility of affine invariance is realized by the fact that general affine deformations can help account for a range of perspective distortions. Specifically we are looking for scalar features F_i's (working on silhouettes) that are invariant to general affine transformations, i.e. $F\{I(u,v)\} = F\{I(x,y)\}$, where

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

A concise and elegant development of affine invariants based on higher order central moments is discussed in [23]. Flusser et al, begin with the assumption that the affine invariant can be expressed in terms of the central moments of the binary image. Drawing from the theory of algebraic invariants, they use two-dimensional moments of the image to derive explicit expressions for independent affine invariants. Here the general two dimensional $(p+q)$'th order central moments of an Image $I(x,y)$ are defined as :

$$\mu_{p,q} = \iint (x-\bar{x})^p (y-\bar{y})^q I(x,y) dx dy$$

The $\bar{x}$ and $\bar{y}$ are the coordinates of the center of gravity of the image. An invariant F is assumed to have the form of a polynomial of the central moments :

$$F = \sum_i k_i \left(\prod_j \mu_{p_j(i),q_j(i)} \right) / \mu_{00}^{z(i)}$$

We include here the final form of the expressions for the first four invariants, and refer the reader for derivations to [23]:

$$\begin{aligned} I_1 &= \left(\mu_{2,0}\mu_{0,2} - \mu_{1,1}^2\right) / \mu_{0,0}^4 \\ I_2 &= \left(\mu_{3,0}^2\mu_{0,3}^2 - 6\mu_{3,0}\mu_{2,1}\mu_{1,2}\mu_{0,3} + 4\mu_{3,0}\mu_{1,2}^2 \right. \\ &\quad \left. + 4\mu_{2,1}\mu_{0,3}^2 - 3\mu_{1,2}^2\mu_{2,1}^2\right) / \mu_{0,0}^{10} \end{aligned}$$

[1] Other solutions such as thresholding the optical flow, or working directly with optical flow magnitude produced almost identical results.

$$
\begin{aligned}
I_3 &= \left(\mu_{2,0}\left(\mu_{2,1}\mu_{0,3} - \mu_{1,2}^2\right)\mu_{0,3}^2 - \mu_{1,1}\left(\mu_{3,0}\mu_{0,3} - \mu_{2,1}\mu_{1,2}\right)\right. \\
&\quad \left. + 3\mu_{0,2}\left(\mu_{3,0}\mu_{1,2} - \mu_{2,1}^2\right)\right)/\mu_{0,0}^7 \\
I_4 &= \left(\mu_{20}^3\mu_{03}^2 - 6\mu_{20}^2\mu_{11}\mu_{12}\mu_{03} - 6\mu_{20}^2\mu_{02}\mu_{21}\mu_{03}\right. \\
&\quad +9\mu_{20}^2\mu_{02}\mu_{12}^2 + 12\mu_{20}\mu_{11}^2\mu_{21}\mu_{03} \\
&\quad +6\mu_{20}\mu_{11}\mu_{02}\mu_{30}\mu_{03} - 18\mu_{20}\mu_{11}\mu_{02}\mu_{21}\mu_{12} \\
&\quad -8\mu_{11}^3\mu_{30}\mu_{03} - 6\mu_{20}\mu_{02}^2\mu_{30}\mu_{12} + 9\mu_{20}\mu_{02}^2\mu_{21}^2 \\
&\quad +12\mu_{11}^2\mu_{02}\mu_{30}\mu_{1}2 - 6\mu_{11}\mu_{02}^2\mu_{30}\mu_{21} \\
&\quad \left. + \mu_{02}^3\mu_{30}^2\right)/\mu_{00}^{11}
\end{aligned} \tag{1}
$$

The idea of using moments on motion regions is not new. In [7] Hu moments are used on a description of the spatial distribution of motion for recognition of activities. However, Hu moments are invariant only under translation, rotation and scaling of the object. By using the moments proposed in [23], we obtain a representation of the moving shape invariant to general affine transformations.

It should be noted that moment invariants are particularly natural for binary silhouettes. Furthermore the invariance to translation eliminates the need for tracking people, body parts or blobs; a common preprocessing scheme in gait modeling.

In this section we outlined a representation that is simple but powerful. We will use this descriptor in the next sections to isolate the dynamics of a gait from its image sequences. We will then discuss how to go from time trajectories of features to dynamical models and cast the gait classification as recognition in the space of linear dynamical systems.

3 Dynamic Modeling with Invariant Moments

We make the assumption that temporal behavior of the invariant moments as the gait evolves in time, can be sufficiently represented as a realization from a second-order stationary stochastic process. This means that the joint statistics between two instants is shift-invariant. This is a restrictive assumption that will allow for modeling of stationary gaits and not for "transient" actions. It is well known that a positive definite covariance sequence with rational spectrum corresponds to an equivalence class of second-order stationary processes. It is then possible to choose as a representative of each class a Gauss-Markov model - that is the output of a linear dynamical system driven by white, zero-mean Gaussian noise - with the given covariance. In other words, we can assume that there exists a positive integer , a process (the "state") with initial condition $x_0 \in \mathbb{R}^n \sim \mathcal{N}(0, P)$ and a symmetric positive semi-definite matrix $\begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \geq 0$ such that $\{y(t)\}$ is the output of the following Gauss-Markov "ARMA" model[2]:

$$
\begin{cases} x(t+1) = Ax(t) + v(t) \quad v(t) \sim \mathcal{N}(0, Q); \;\; x(0) = x_0 \\ y(t) = Cx(t) + w(t); \quad w(t) \sim \mathcal{N}(0, R) \end{cases} \tag{2}
$$

for some matrices $A \in \mathbb{R}^{n \times n}$ and $C \in \mathbb{R}^{m \times n}$.

The first observation concerning the model (2) is that the choice of matrices A, C, Q, R, S is not unique, in the sense that there are infinitely many models that give rise to exactly the same measured covariance sequence starting from suitable initial conditions. The first source of non-uniqueness has to do with the choice of basis for the state space: one can substitute A with TAT^{-1}, C with CT^{-1}, Q with TQT^T, S with TS, and choose the initial condition Tx_0, where $T \in \mathcal{GL}(n)$ is any stable $n \times n$ matrix and obtain the same output covariance sequence; indeed, one also obtains the same output realization. The second source of non-uniqueness has to do with issues in spectral factorization that are beyond the scope of this paper [28]. Suffices for our purpose to say that one can transform the model (2) into a particular form – the so-called "innovation representation" – that is unique. In order to be able to identify a unique model of the type (2) from a sample path $y(t)$, it is therefore necessary to choose a representative of each equivalence class (i.e. a basis of the state-space): such a representative is called a *canonical model realization* (or simply canonical realization). It is canonical in the sense that it does not depend on the choice of the state space (because it has been fixed).

While there are many possible choices of canonical realizations (see for instance [29]), we are interested in one that is "tailored" to the data, in the sense of having a diagonal state covariance. Such a model realization is called *balanced* [30]. The

[2] ARMA stands for autoregressive moving average.

problem of going from data to models then be formulated as follows: *given* measurements of a sample path of the process: $y(1), \ldots, y(\tau);\ \tau >> n$, estimate $\hat{A}, \hat{C}, \hat{R}, \hat{Q}$, a canonical realization of the process $\{y(t)\}$. Ideally, we would want the maximum likelihood solution from the finite sample, that is the argument of

$$\max_{A,C,Q,R} p(y(1), \ldots, y(\tau)|A, C, Q, R). \tag{3}$$

The closed-form asymptotically optimal solution to this problem has been derived in [31]. From this point on, therefore, we will assume that we have available – for each sample sequence – a model in the form $\{A, C, Q, R\}$. While the state transition A and the output transition C are an intrinsic characteristic of the model, the input and output noise covariances Q and R are not significant for the purpose of recognition (we want to be able to recognize trajectories measured up to different levels of noise as the same). Therefore, from this point on we will concentrate our attention on the matrices A and C that describe a gait.

4 Recognizing Gaits

Models, learned from data as described in the previous section, do not live on a linear space. While the matrix A is only constrained to be stable (eigenvalues within the unit circle), the matrix C has non-trivial geometric structure for its columns form an orthogonal set. The set of n orthogonal vectors in $\mathbb{R}^m$ is a differentiable manifold called "Stiefel manifold".

Because of the highly curved structure of this space, state-of-the-art classification algorithms applied on the model parameters fail to produce satisfactory results. In particular, we tested an efficient implementation [25] of the Support Vector Machines classifier [24] on the vectors obtained by stacking the elements of the matrices A and C. With this approach discrimination was not possible even in the simple case of only two classes of gaits.

4.1 Distance Between Models

As proposed in [4], a natural solution for the recognition problem in this case is provided by endowing the space of models with a metric structure. In the literature of system identification and signal processing, the problem of defining a metric in the space of linear dynamical systems is an active area of research [26, 27]. A common distance that is widely accepted in system identification for comparing ARMA models is based on the so-called subspace angles [31].

Given a model M specified by the matrices (A, C), the infinite observability matrix $O(M)$ is defined as:

$$O(M) = \begin{bmatrix} C^T & A^T C^T & A^{2T} C^T & \cdots \end{bmatrix} \in \mathbb{R}^{\infty \times n}$$

The matrix $O(M)$ spans an n-dimensional subspace of $\mathbb{R}^\infty$. To compare two models M_1 and M_2, the basic idea is to compare "angles" between the two observability subspaces of M_1 and M_2. There are many equivalent ways to define subspace angles. Given a matrix H with its columns spanning an n-dimensional subspace, let Q_H denote the orthonormal matrix which spans the same subspace as H. Given two matrices H_1, H_2, we denote the n ordered singular values of the matrix $Q_{H_1}^T Q_{H_2} \in \mathbb{R}^{n \times n}$ to be $cos^2(\theta_1), ..., cos^2(\theta_n)$. Then the principal angles between subspaces spanned by H_1 and H_2 are denoted by the n-tuple:

$$H_1 \wedge H_2 = (\theta_1, \theta_2, ..., \theta_n) \quad , \theta_i \geq \theta_{i+1} \geq 0.$$

Based on these angles, two distances can be defined:

$$d_M^2 = -\ln \prod_i \cos^2(\theta_i), \quad d_F = \theta_1. \tag{4}$$

The first distance is an extension of Martin distance defined for SISO systems [27], the second is the Finsler distance according to Weinstein [32]. Roughly speaking, the difference between these two distances is that d_M^2 is an L^2-norm but d_F is an L^∞-norm between linear systems.

Once a metric in the space of models is available, standard grouping techniques such as k-means clustering can be successfully employed for recognition.

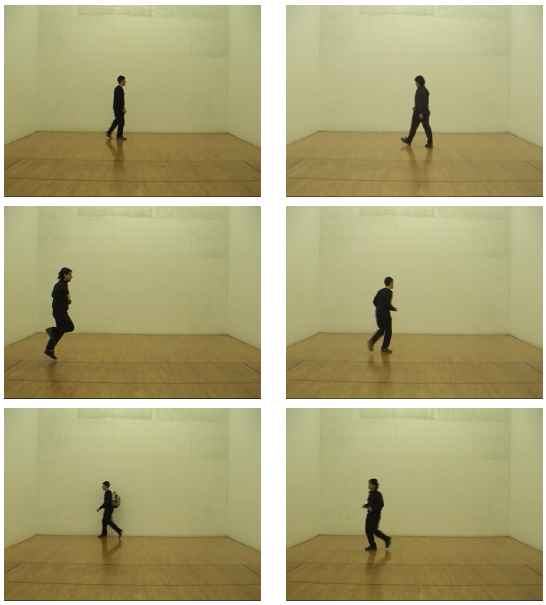

Figure 1: *Sample frames from the dataset of the gaits: walking, running, jumping and limping.*

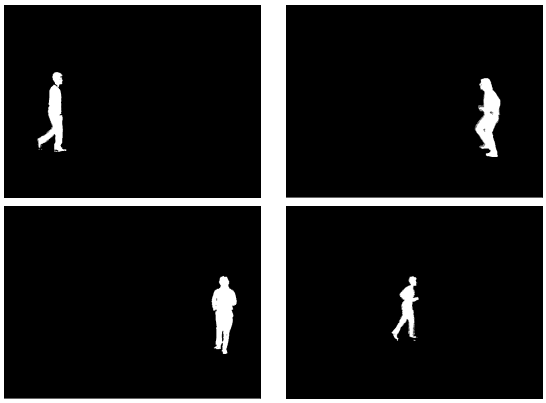

Figure 2: *Sample silhouettes extracted by background subtraction.*

5 Experiments and Results

Our gait dataset consists of short clips of walking (with and without a backpack), limping, running and jumping performed by two subjects, for a total of 81 sequences. In Table 4 we show a more detailed description of the exerimental data, and in Figure 1 sample frames from the video sequences.

Given a gait sequence, for each frame we used backround subtraction to extract a sihouette of the moving body, and computed the affine invariant moments on this binary image. Figure 2 shows sample output of the background subtraction, and in Figure 3 the trajectories of the moments for some sequences in the dataset are plotted. From the experiments, we noticed that moments of order higher than 4 are too sensible to noise and negatively affect the results. Also, the four moments (1) have different scales and need be normalized to form the feature vector $y(t)$. The values of the scale factors were found empirically by matching the mean energy of the moments.

For each sequence of moment trajectories $y(t)$ we have identified a dynamical model of orders $n = 1$ to 4. For identifying the model we used the implementation of the N4SID algorithm [33] in the Matlab System Identification Toolbox. Since our models are zero-mean, we subtract the mean from the data before the learning step.

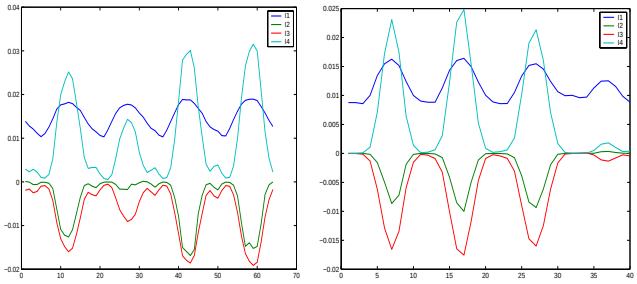

Figure 3: *Plots of affine invariant moments computed on the binary silhouettes: on the left moments from a walking sequence of subject A, on the right moments from a running sequence of subject B.*

Gait	Number of Sequences	
	Subject A	Subject B
Walking	28, 8 with backpack	13
Running	9	11
Jumping	9	4
Limping	7	0

Figure 4: *Description of the gait dataset: 4 gait classes performed by 2 persons for a total of 81 sequences, details as above.*

We then computed the mutual distance between each model by calculating the distances between observability subspaces: Finsler distance d_F and our generalization of Martin distance d_M, as defined in (4). These two distances gave similar results, with an advantage for the latter one. In figure 5 we show the pairwise distance between models of sequences in the dataset, with highlighted the two nearest neighbors.

As the result show, the dynamics of the invariant is able to distinguish between different sytles of gaits while preserving the invariance with respect to apperance and geometric factors. Discrimination fails only when comparing sequences of walking and limping, due to the high similarity of these two classes of motion.

6 Conclusions

We introduced a method to incorporate both appearance and geometric affine invariances to represent gait sequences. We discussed how to extract such invariant representations from images and model their temporal behavior. We then developed a frame work where invariant representations could be modelled and compared in the space of linear dynamical systems. We presented results using over 80 sequences of gait samples corresponding to various different viewing angles and gait types. We were able to cluster and isolate sequences corresponding to same or similar gaits regardless of the subject or geometric viewing factors. The power of affine invariant representation and richness of dynamical systems for modeling temporal structure of gaits is present in the results.

References

[1] Y. Song, X. Feng, and P. Perona, Towards detection of human motion. In *Proc. of CVPR*, pages 810-817, 2000.

[2] C. Sminchisescu and B. Triggs. Kinematic Jump Processes for Monocular Human Tracking. In *Proc. of CVPR*, 2003

[3] D. M. Gavrila. The visual analysis of human movement: A survey. In *Computer Vision and Image Understanding*, volume 73, pages 82–98, 1999.

[4] Omitted during the reviewing process

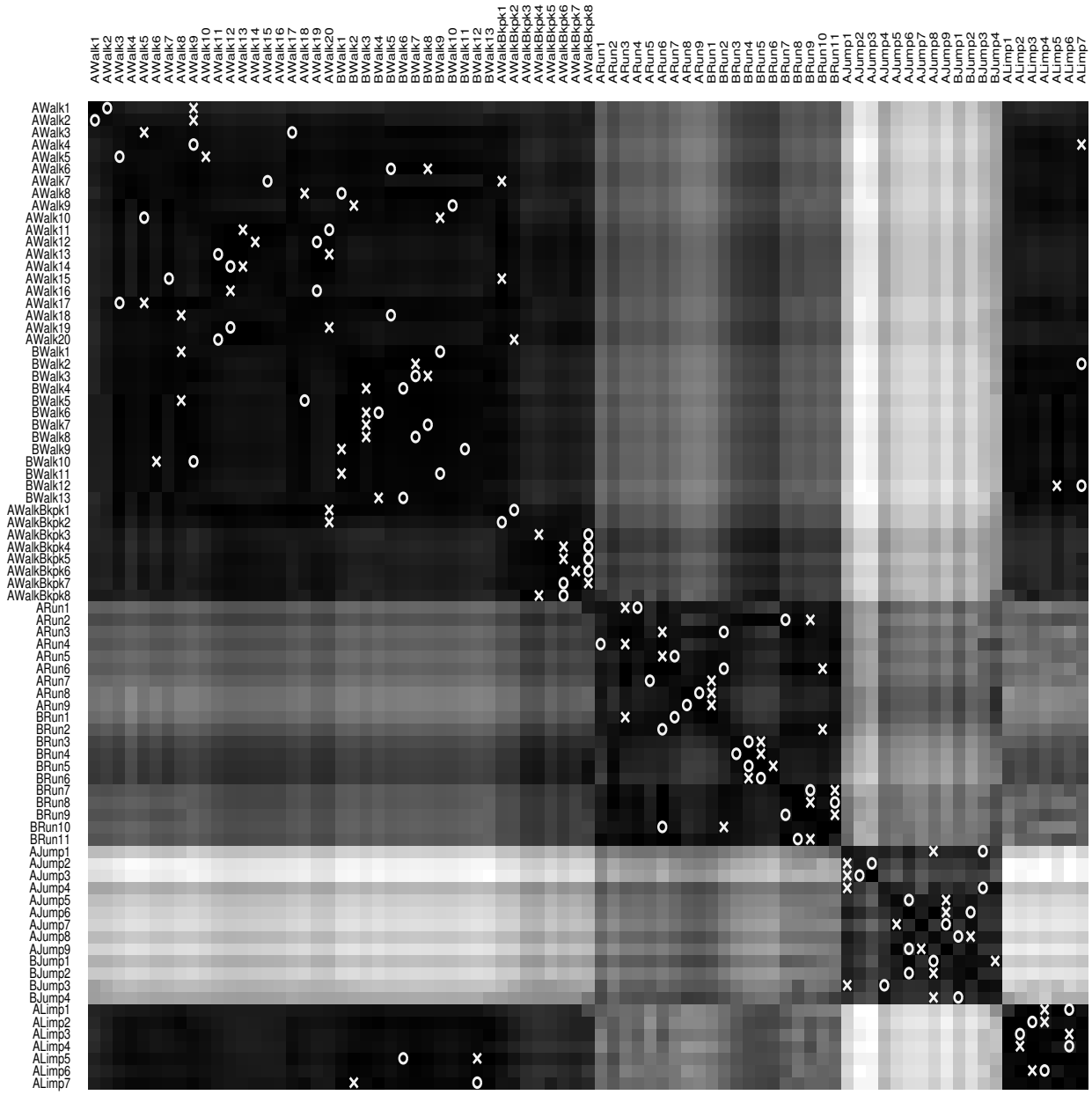

Figure 5: *Confusion matrix depicting the pairwise distance between models. Each row/column represents a sequence, and sequences of the same class are grouped in blocks. Dark indicates a small distance, light a large distance. For each row, a circle indicates the nearest neighbor and a cross identifies the second nearest neighbor. The last block corresponds to the limping gaits. As we could expect, there are some misclassifications between limping and walking, since these gaits are very close. For the other classes of gait, the block diagonal structure of the matrix clearly indicates the effectiveness of our approach.*

[5] T.Starner and A. Pentland. Real-time american sign language recognition from video using hmm. In *Proc. of ISCV 95*, volume 29, pages 213–244, 1997.

[6] A. D. Wilson and A. F. Bobick. Parametric hidden markov models for gesture recognition. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, volume 21(9), pages 884–900, Sept. 1999.

[7] A. F. Bobick. and J. W. Davis The recognition of human movement using temporal templates. In *IEEE Trans. PAMI*, 23(3):257-267, 2001.

[8] J. J. Little and J. E. Boyd. Recognizing people by their gait: the shape of motion. 1996.

[9] C. R. Wren and A. P. Pentland. Dynamic models of human motion. In *Proceedings of FG'98*, Nara, Japan, April 1998.

[10] C. Bregler. Learning and Recognizing Human Dynamics in Video Sequences. In *Proc of CVPR*, pp. 568-574, 1997

[11] C. Bregler and J. Malik Tracking People with Twists and Exponential Maps In *Proc. of CVPR*, 1998

[12] V. Pavlovic and J. Rehg and J. MacCormick. Impact of Dynamic Model Learning on Classification of Human Motion In *Proc. of International Conference on Computer Vision and Pattern Recognition*, 2000.

[13] A. A. Efros, A. C. Berg, G. Mori and J. Malik Recognizing Action at a Distance In *Proc. of International Conference on Computer Vision*, 2003.

[14] R. Cutler and L. Davis Robust real-time periodic motion detection, analysis, and applications. In *IEEE Trans. PAMI*, 22(8), August 2000.

[15] L. Zelnik-Manor and M. Irani Event-based video analysis. In *Proc of CVPR*, 2001.

[16] H. Sidenbladh, M. J. Black and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *Proc. European Conference on Computer Vision*, vol 1, pp 784-800, 2002

[17] M. E. Brand and A. Hertzmann. Style Machines. ACM SIGGRAPH, pps 183-192, July 2000

[18] B. North and A. Blake and M. Isard and J. Rittscher. Learning and classification of complex dynamics. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, volume 22(9), pages 1016-34, 2000.

[19] M. J. Black and A. D. Jepson. A Probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. In *Proc. of European Conference on Computer Vision*, volume 1, pages 909-24, 1998.

[20] R. Polana and R. C. Nelson Detection and recognition of periodic, non-rigid motion. In *Int. Journal of Computer Vision*, 23(3):261-282, 1997.

[21] H. A. Rowley and J. M. Rehg Analyzing articulated motion using expectation-maximization. In *Proc. CVPR*, 1997

[22] M. K. Hu Visual pattern recognition by moment invariants. In *IRE Trans. Inf. Theory*, IT-8, 179-187, 1962

[23] J. Flusser and T. Suk Pattern Recognition by Affine Moment Invariants In *Pattern Recognition*, vol 26, No. 1, pp. 167-174, 1993

[24] V. N. Vapnik The Nature of Statistical Learning Theory. Springer, 1995.

[25] T. Joachims Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schoelkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[26] K. De Coch and B. De Moor. Subspace angles and distances between arma models. *Proc. of the Intl. Symp. of Math. Theory of Networks and Systems*, 2000.

[27] R. Martin. A metric for arma processes. *IEEE Trans. on Signal Processing*, 48(4):1164–1170, 2000.

[28] L. Ljung. *System Identification: theory for the user*. Prentice Hall, 1987.

[29] T. Kailath. *Linear Systems*. Prentice Hall, 1980.

[30] K Arun and S. Y. Kung. Balanced approximation of stochastic systems. *SIAM Journal of Matrix Analysis and Applications*, 11(1):42–68, 1990.

[31] P. Van Overschee and B. De Moor. Subspace algorithms for the stochastic identification problem. *Automatica*, 29:649–660, 1993.

[32] A. Weinstein Almost invariant submanifolds for compact group actions Berkeley CPAM Preprint Series n.768, 1999

[33] P. V. Overschee and B. De Moor N4SID: Subspace Algorithms for the Identification of Combined Deterministic-Stochastic Systems. In *Automatica, Special Issue on Statistical Signal Processing and Control*, Vol. 30, No.1, 1994, pp. 75-93

[34] G. Johansson. Visual perception of biological motion and a model for its analysis. Perception and Psychophysics, 14(2):201–211, 1973.