Recognition of Human Gaits

Alessandro Bissacco* Alessandro Chiuso† Yi Ma‡ Stefano Soatto*

*Computer Science
University of California, Los Angeles
Los Angeles - CA 90095

†Dipartimento di Elettronica e Informatica
Università di Padova
35131 Padova, Italy

‡Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana - IL 61801

Abstract

We pose the problem of recognizing different types of human gait in the space of dynamical systems where each gait is represented. Established techniques are employed to track a kinematic model of a human body in motion, and the trajectories of the parameters are used to learn a representation of a dynamical system, which defines a gait. Various types of distance between models are then computed. These computations are non trivial due to the fact that, even for the case of linear systems, the space of canonical realizations is not linear.

1. Introduction

How do we recognize a person walking from one jumping, running, hopping or dancing, independent on the person and her pose? In this paper we address this problem in three stages by first tracking a parametric skeletal model of the person moving, then learning a model that captures the dynamics of the model parameters, and finally posing the recognition problem in the space of dynamical systems learned from data. The first stage renders the system invariant to photometric factors (e.g. illumination, clothing etc.), while the second stage guarantees invariance to geometric factors such as the distance and pose of the camera, the length of the person's limbs etc. In this context, each gait is represented by a dynamical system identified from a time series.

While we borrow the first stage straight from the literature of computer vision, subsequent stages require some attention. Learning a dynamical model of the joint trajectories must be done in a canonical way to guarantee that a particular dataset corresponds to one and only one model. This can be done following well established results in the literature of system identification, at least for the simplest case of linear systems. Performing recognition in the space of models entails computing distances and probability distributions on manifolds, since the set of dynamical models in canonical form is *not* a linear space (even if the model itself is linear!) and therefore computing distances naively can lead one to conclude that very similar models are dramatically different. Measuring distances between dynamical models is an open problem even for the case of linear multiple-input, multiple-output (MIMO) systems. Endowing the space of models with a full-fledged probabilistic structure is even more of a challenge.

Our starting point is a collection of trajectories of joint positions and/or joint angles for an articulated body. We extract those from images using an algorithm similar to that of Bregler and Malik [6], manually initialized. Although manual initialization is not our ideal choice, performing automatic initialization is a research program in its own right, and is therefore not addressed in this paper. The emphasis of our work is *not* tracking. Therefore, we consider the output of any tracking module as the input of our algorithm that estimates a dynamical model of the geometric feature trajectories.

The literature on modeling and recognition of human motion is sizeable and growing (see [10] for a survey). A common approach consists of extracting low-level features by local spatio-temporal filtering on the images and using hidden Markov models (HMMs) on the collection of sequences of points thus obtained for recognition and classification tasks [24, 25]. In [25], parametric HMMs are introduced for recognizing gestures that exhibit dependence on a set of parameters, and in [5] coupled HMMs are used for modeling interactions of two mobile parts. In [17, 1] Bayesian Networks are used for recognition tasks. Local representation of motion based on optical flow has been exploited in [14, 15], and view-based methods are proposed in

[4, 2, 12]. Other approaches are based on principal component analysis [28], parameterization of the motion on joint angles [7] and snake fitting [19]. Estimation of motion from stereo [27] and multiple view systems [11] has also been investigated. In [6] a mixed-state statistical model for the representation of motion has been been proposed. In this switching linear dynamic model a stochastic finite-state automaton at the highest level switches between local linear Gaussian models. Estimation and recognition is performed with expectation-maximization (EM) approaches using particle filters [20, 3] or structured variational inference techniques [23].

Our models are discrete-time, continuous-state dynamical systems, and the action is coded in the dynamical model (i.e. the system parameters). We use closed-form algorithms [22] rather than EM as customary, to perform learning. Since the space of model is non-linear, computing a distance between models is non trivial. We draw on the literature of system identification and signal processing, where the problem is an active area of research [8, 18]. We propose different methods that, on the admittedly limited dataset we have tried them on, give encouraging results.

2. Preliminaries

2.1. From images to skeletons

In this section we briefly describe the algorithm we use to extract joint trajectories from pictorial image sequences. The emphasis of this work is on the recognition of gaits. Therefore, any algorithm for the estimation of joint trajectory – reviewed in Section 1 – can be used as a front-end.

In particular, we have implemented a variant of [6], where a human skeleton is represented as a kinematic chain supporting ellipsoidal texture patches. The parameters of the chain are set by hand by clicking on the desired joints in the first image of a sequence. Each link is then represented by an ellipse whose major axis equals the length of the link and whose minor axis is also set by hand. At each tracking step, an EM iteration is performed where the joint parameters are estimated for a given support region, followed by an update of the support region based on a local measure of similarity among corresponding regions in adjacent images.

The result of this algorithm – or of any other similar algorithm – is a sequence of joint positions, which we call $y(t), t = 1 \dots \tau$. These are used to compute a dynamical model of the joint evolution, as we describe next.

2.2. From joint angle trajectories to dynamical models

We start from the assumption that a sequence of joint angle trajectories $y(t),\ t = 1 \dots \tau$ is a realization from a second-order stationary stochastic process. This means that the joint statistics between two instants is shift-invariant. This is a severely restrictive assumption that is only meaningful for stationary gaits but 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 [16]. 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 n, a process $\{x(t)\}$ (the "state") with initial condition $x_0 \in \mathbb{R}^n \sim \mathcal{N}(0, P)$ and a symmetric positive semi-definite matrix $\left[\begin{array}{cc} Q & S \\ S^T & R \end{array}\right] \geq 0$ such that $\{y(t)\}$ is the output of the following Gauss-Markov "ARMA" model[1]:

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

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

The first observation concerning the model (1) is that the choice of matrices A, C, Q, R, S is not unique. 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. The second source of non-uniqueness has to do with issues in spectral factorization that are beyond the scope of this paper [16]. Suffices for our purpose to say that one can transform the model (1) into a particular form – the so-called "innovation representation" – that is unique. Such a representation is canonical in the sense that it does not depend on the choice of the state space (because it has been fixed).

The problem of going from data to models then be formulated as follows: *given* measurements of a sample path of the process: $y(1), \dots, y(\tau);\ \tau >> n$, 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

$$\max_{A,C,Q,R} p(y(1), \dots, y(\tau) | A, C, Q, R). \tag{2}$$

The closed-form asymptotically optimal solution to this problem has been derived in [22]. 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

[1] ARMA stands for autoregressive moving average.

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.

2.3 Distance between models

A common distance that is widely accepted in system identification for comparing ARMA models is based on the so-called subspace angles [22]. Given a model M specified by the pair (A, C) as above, one may define the associated *infinite observability matrix*,

$$\mathcal{O}(M) = [C^T \; A^T C^T \; A^{2T} C^T \; \cdots]^T \in \mathbb{R}^{\infty \times n}. \quad (3)$$

One may view the matrix $\mathcal{O}(M)$ as an n-dimensional subspace of $\mathbb{R}^\infty$ that is spanned by its n columns. To compare two models M_1 and M_2, the basic idea is then to compare "angles" between the two observability subspaces of M_1 and M_2. A canonical notion of angle between two subspaces is given by the so-called *subspace angles*, also known as *principal angles*. There are many (algebraically or geometrically) equivalent ways to define subspaces angles (see [22, 13, 26]). Here we only introduce one which is conceptually simple: 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 (which can be found through the Gram-Schmidt orthogonalization). 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 $\cos^2(\theta_1), \ldots, \cos^2(\theta_n)$. Then the principle angles between subspaces spanned by H_1 and H_2 are denoted by the n-tuple:

$$H_1 \wedge H_2 = (\theta_1, \theta_2, \ldots, \theta_n), \quad \theta_i \geq \theta_{i+1} \geq 0. \quad (4)$$

Based on these angles, two distances can be defined:

$$d_M^2 = -\ln \prod_i \cos^2(\theta_i), \quad d_F = \theta_1. \quad (5)$$

The first distance d_M^2 follows the definition of Martin [18] and the second is according to Weinstein [26]. In the case of the Martin distance d_M^2, for minimum phase single-input single-output (SISO) systems, it is equivalent to the norm deduced from a natural metric on the cepstrum of the system auto-correlation function [18], and as shown in [8] this distance has a closed-form formula in terms of the systems' poles and zeros. However, for MIMO systems, it is not even guaranteed that the quantity d_M^2 be non-negative. The distance d_F, taking the largest principal angle, is always non-negative and geometrically it is the Finsler distance between the two subspaces viewed as two elements in the Grassman manifold $G(\infty, n)$ [26]. 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.

3. Recognizing gaits

As we have articulated in the previous section, a gait is represented by a linear dynamical system and described by the matrices A, C that live in the space $\mathcal{GL}(n) \times V(m, n)$. This space has a non-trivial curvature structure that must be taken into account when doing comparisons between models.

One issue that we have not elaborated on is the choice of the model order n. This is performed empirically according to the measured value of the canonical correlations, as discussed in the experimental section 4.

¿From a pattern recognition viewpoint, constructing a probability density is not necessary to solve problems such as "clustering" or "grouping". For instance the k-nearest neighbor algorithm only requires a distance to be implemented, and it is therefore easily extended to Stiefel manifolds under the notion of distance that we have defined. Suppose a set of samples $C_1, C_2, \ldots$ is given, where each sample is labeled as belonging to one of c classes λ_j. Given a new sample C, the label λ_m is chosen by taking a vote among the k nearest samples. That is, λ_m is selected if the majority of the k nearest neighbors have label λ_m, which happens with probability

$$\sum_{i=(k+1)/2}^{k} \left(\begin{array}{c} k \\ i \end{array} \right) P(\lambda_m|C)^i (1 - P(\lambda_m|C))^{k-i}. \quad (6)$$

It can be shown [9] that if k is odd the large-sample 2-class error rate is bounded above by the smallest concave function of P^* – the optimal error rate – greater than

$$\sum_{i=0}^{(k-1)/2} \left(\begin{array}{c} k \\ i \end{array} \right) \left({P^*}^{i+1} (1 - P^*)^{ki} + {P^*}^{k-i} (1 - P^*)^{i+1} \right). \quad (7)$$

Note that the analysis holds for k fixed as $n \to \infty$, and that the rule approaches the minimum error rate for $k \to \infty$. For small samples, there are no known results except negative counter-examples that show that an arbitrarily bad error rate can be achieved.

4. Experiments

In this section we describe preliminary experiments in recognizing different types of gait. We have collected several sequences of humans walking, running, dancing, jumping etc., 10 per each gait. On each sequence, we have then defined a reduced kinematic model (half of the skeleton along a sagittal section, since the other half follows by symmetry), and considered the time trajectory of the projection of 4 joints onto the image plane: shoulder, elbow, hip and knee. For each gait, we have changed the viewing position, distance and subject.

For each sequence of joint trajectories we have identified a dynamical model of orders $n = 1$ to 4. For identifying the model we used the implementation of the N4SID algorithm [21] in the Matlab System Identification Toolbox. Since our models are zero-mean, we subtract the mean from the data before the learning step.

We have then computed the mutual distance between each model in a number of ways. First we have computed the "naive" distance (the 2-norm of the difference between corresponding system matrices, without taking the geometry of the subspace into account) and the geodesic distance between models. Not surprisingly, these led to quite disappointing results. Then we have computed two distances between observability subspaces - indeed taking into account the geometry of the subspace: the Finsler distance [26] and a generalization of the Martin distance, defined in [18] for SISO models. We computed the principal angles between observability subspaces using an algorithm similar to the one proposed in [8], extended to MIMO systems under the assumption of full-rank innovation models. Then we have calculated the Finsler distance d_F and the Martin distance d_M^2 as defined in (5). These two distances gave similar results, with an advantage for the latter one. The Matlab code for computing the subspace angles is reported in Figure 4. To the distance between learned zero-mean models we added the norm of the difference between the the means of the joint configurations, weighted by a scale factor whose value was set empirically.

For the purpose of illustration, we show the results of the most challenging experiment with our current dataset, corresponding to three classes of motions that result in similar gaits: walking, running and going up and down a staircase. Notice that these three gaits are quite similar to each other (as opposed, say, to dancing or jumping), and yet the algorithm proposed is capable of distinguishing between them with high probability. In Figure 1 we show sample frames from the training datasets. Figure 2 shows the pairwise distance between each model in the dataset. As it can be seen, similar gaits result in smaller distances, with a few outliers. Although this is a very restricted database, it suffices to test our hypothesis.

We have now chosen a few sample sequences for each category as a test sequence. For each of the sequences we have estimated a model by first pre-processing the sequence (after manual initialization) using the ideas described in [6] to extract joint coordinates, and finally compared the models using a nearest neighbor criterion. A sample frame from the test sequence is shown in Figure 3, while the first two corresponding nearest neighbors are shown to the right. Although this dataset is quite small, the discriminating power of the model as a representation of the dynamic sequence is visible. Acquiring an extensive dataset of gaits under a variety of viewing conditions is part of our future research agenda.

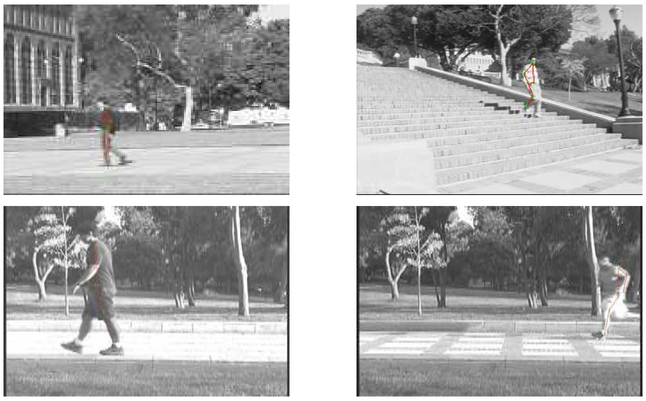

Figure 1. *Sample frames from the dataset: waking, running and walking a staircase.*

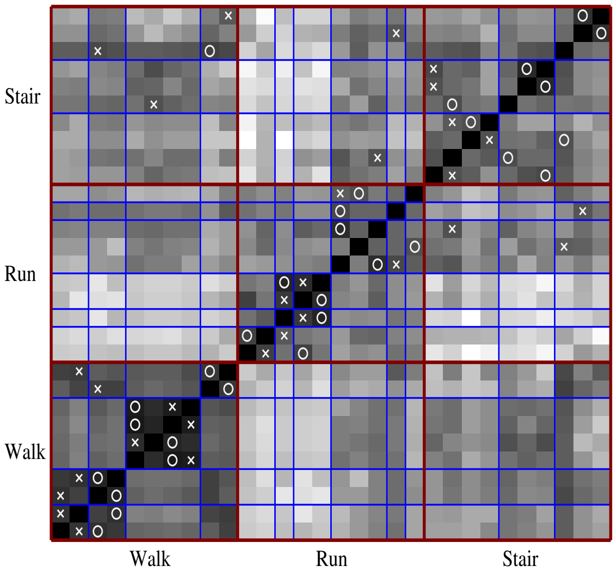

Figure 2. *The pairwise distance between each sequence in the dataset is displayed in this plot. Each row/column of a matrix represents a sequence, and sequences corresponding to similar gaits are grouped in block rows/columns. Dark indicates a small distance, light a large distance. The minimum distance is of course along the diagonal; for each row the next closest sequence is indicated by a circle, while the second nearest is indicated by a cross.*

```
function theta = subspace_angles(A1,K1,C1,A2,K2,C2)

n = size(A1,1);
m = size(C1,1);
A = [ A1 zeros(n,3*n); zeros(n) A2-K2*C2 zeros(n,2*n);
     zeros(n,2*n) A2 zeros(n); zeros(n,3*n) A1-K1*C1];
C = [ C1 -C2 C2 -C1 ];
Q = dlyap(A',C'*C);
E = eig([zeros(2*n) pinv(Q(1:2*n,1:2*n))*Q(1:2*n,2*n+1:4*n);
         pinv(Q(2*n+1:4*n,2*n+1:4*n))*Q(2*n+1:4*n,1:2*n)
         zeros(2*n)]);
E = max(-ones(size(E)),E);
E = min(ones(size(E)),E);
theta = acos(E(1:2*n));
```

Figure 4. *Matlab code for computing subspace angles between innovation models (`dlyap` is a function of the System Identification Toolbox)*

References

[1] J. Binder, D. Koeller, S. Russell, and K. Kanazawa. Adaptive probabilistic networks with hidden variables. In *Machine Learning*, volume 29, pages 213–244, 1997.

[2] M. J. Black. Eigentracking: robust matching and tracking of articulated objects. 1996.

[3] 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.

[4] A. F. Bobick. Appearance-based representation of action. 1996.

[5] M. Brand, N. Oliver, and A. Pentland. Coupled hmm for complex action recognition. In *Proc. of Conference on Computer Vision and Pattern Recognition*, volume 29, pages 213–244, 1997.

[6] C. Bregler. Learning and recognizing human dynamics in video sequences. In *Proc. of the Conference on Computer Vision and Pattern Recognition*, pages 568–574, 1997.

[7] L. Campbell and A. Bobick. Recognition of Human Body Motion Using Phase Space Constraints. In *Proc. International Conference on Computer Vision and Pattern Recognition*, pages 624-630, Cambridge MA, 1995.

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

[9] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Trans. on Information Theory*, 13:21–27, 1967.

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

[11] D. M. Gavrila and L. S. Davis. Tracking of humans in action: a 3-d model-based approach. 1996.

[12] M. A. Giese and T. Poggio. Morphable models for the analysis and synthesis of complex motion patterns. In *International Journal of Computer Vision*, volume 38(1), pages 1264–1274, 2000.

[13] G. Golub and H. Zhu. The canonical correlations of matrix pairs and their numerical computation. *Linear Algebra for Signal Processing*, the IMA volumes in mathematics and its applications, volume 69, pages 27-49, 1992.

[14] J. Hoey and J. J. Little. Representation and recognition of complex human motion. In *Proc. of the Conference on Computer Vision and Pattern Recognition*, volume 1, pages 752–759, 2000.

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

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

[17] A. Madabhushi and J. K. Aggarwal. A bayesian approach to human activity recognition. In *Proc. of the 2nd International Workshop on Visual Surveillance*, pages 25–30, June 1999.

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

[19] A. A. Niyogi. Analyzing and recognizing walking figures in xyt. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, pages 469–474, Seattle, June, 1994.

[20] 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.

[21] P. Van 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.

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

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

[24] 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.

[25] 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.

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

[27] C. Wren. Dynamic models of human motion. 1998.

[28] Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. In *Computer Vision and Image Understanding*, volume 73(2), pages 232–247, 1999.

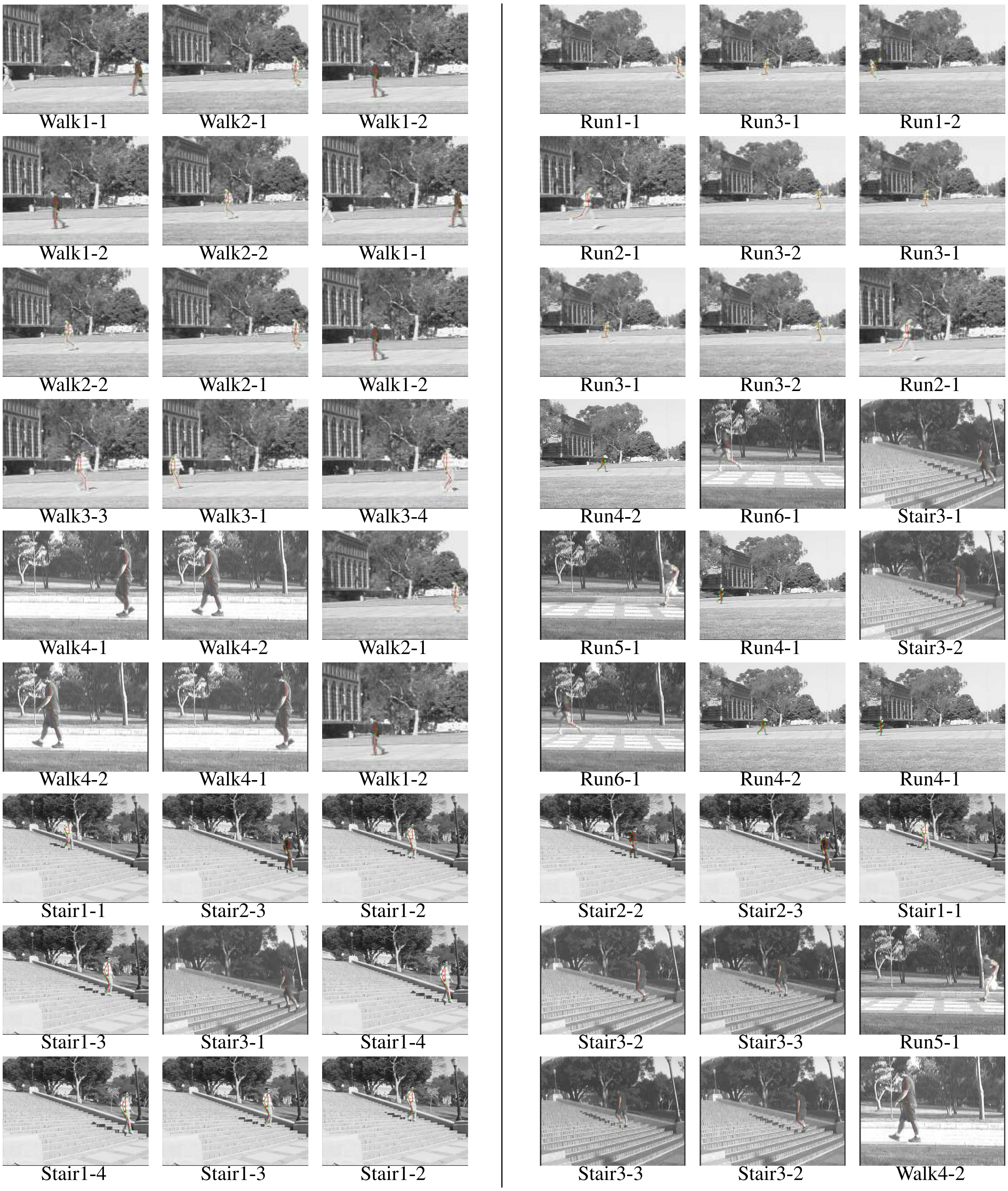

Figure 3. *For each gait we have chosen a few sample sequences (left) and computed the distance to every other sequence in the dataset. The closest sequence is shown in the central column, while the second nearest is shown in the right column. With a few exceptions, the nearest neighbor belongs to the same gait as the test sequence. Notice that all gaits are quite similar; similar experiments performed on much more diverse gaits such as jumping or dancing return correct classifications. More extensive experimental evaluations are forthcoming.*