

Quantification and classification of locomotion patterns by spatio-temporal morphable models

M.A. Giese and T. Poggio

Center for Biological and Computational Learning
Massachusetts Institute of Technology, E25-206 / 218
Cambridge, MA 02142, USA
Tel.: 617 253 0549 / 5230, Fax: 617 253 2964

E-mail: giese@mit.edu
tp@ai.mit.edu

Third IEEE Workshop on Visual Surveillance,
July 1, 2000, Dublin, Ireland

Quantification and classification of locomotion patterns by spatio-temporal morphable models

M. A. Giese and T. Poggio
Center for Biological and Computational Learning
Massachusetts Institute of Technology, E25-206 / 218
Cambridge, MA 02142, USA

Abstract

Morphable models have been applied successfully in the context of computer vision and computer graphics for the representation of classes of stationary images. In this paper, we develop a similar technique for the representation of classes of complex movements that we call space-time morphable models. This technique permits to approximate new complex movement patterns by linear combinations of few learned prototypical example patterns. The weights of the linear combination provide a low-dimensional description of the patterns that can be exploited for the classification of the underlying actions, and also for the estimation of continuous parameters that quantify characteristic properties of the movement. (Examples are the direction of locomotion and the style with which a certain movement is executed.) We demonstrate the applicability of the technique for the classification and quantification of properties of locomotion patterns. Several possible applications of space-time morphable models in the context computer vision and surveillance are discussed.

Keywords: recognition, prototype, linear superposition, learning, correspondence

1 Introduction

The analysis of complex movement patterns is an important problem for automatic surveillance systems, but also for other computer vision applications, like multi-media interfaces. The recognition of actions and gestures has been a popular theme in computer vision for the last few years. Many known methods permit a classification of different complex movement patterns (see [9] for a review). Different approaches have been proposed, like the direct extraction of spatio-temporal features from image sequences (e.g. [17, 6, 7]) or the combination of 2D or 3D models for the human body or face with predictive filtering techniques [18, 7, 23] or Hidden Markov Models (e.g. [20]). A broad

literature exists also on the tracking of nonrigidly moving objects. Most of these methods are model-based, and use either predesigned models, or they construct such models by learning of example images of the tracked object (e.g. [3, 1, 4]).

Even though many methods exist that allow a classification of visually observed action patterns, not much work has investigated the recognition of more subtle aspects of observed movements. For surveillance purposes, it may for instance be interesting not only to recognize that a person is walking, but also to estimate in which direction the person is walking. Similarly, it might be interesting to quantify the style of movements, and to estimate e.g. if a person is walking very dynamically, or in a rather hesitating way. The method that is presented in this paper permits to determine such information from video tracking data. It is based on a technique that we call *spatio-temporal morphable models* which permits to approximate complex spatio-temporal patterns by linear combinations of learned example patterns which are defined by feature trajectories. Linear combinations are defined by spatio-temporal morphing between the learned examples. Based on the linear weights that determine the contributions of the examples to the linear combination, complex movement patterns can be classified and continuous parameters that characterize the movements can be estimated.

The paper is structured as follows: We first develop the concept of spatio-temporal morphable models (section 3) by generalizing morphable models for stationary image classes (reviewed in section 2). The central theoretical problem for the construction of morphable models is to establish correspondence between spatio-temporal patterns. Section 4 sketches an algorithm for the calculation of such correspondences. In section 5 we discuss how morphable models can be used for the classification of movements and for the estimation of continuous parameters that characterize their properties. Section 6 describes the tracking data that we used to test the method. The results of our evaluation are presented in section 7. Finally, we discuss possible applications and extensions of our method in section 8.

2 Morphable models for stationary images

Morphable models have been originally introduced for the representation of classes of stationary images. It was shown that appropriately defined linear combinations of few prototypical images of a three-dimensional objects can approximate another view of the object very accurately. This method has been used for the synthesis of new virtual images from example images (e.g. [13, 22, 21]). Meanwhile, morphable models have been used in a broad spectrum of technical applications. Examples are the recognition and synthesis of face images [2, 22], the synthesis of new facial expressions [5], or the simulation of talking faces [8].

Morphable models approximate new images by linear combinations of learned prototypical images. If one defines "linear combinations" by morphing between the prototypical examples one obtains a learning-based representation with very efficient generalization properties. This makes it possible to represent classes of natural images on the basis of a small number of learned prototypical images. This distinguishes morphable models from other methods, like the linear combination of principle components of the brightness distributions ("eigen faces") [19], which require typically a larger number of prototypes to achieve accurate approximation.

How can the linear combination of a small number of prototypical images be defined in a way that ensures good generalization? If the linear combination of two images is defined by linearly combining the brightness values, like in the case of "eigen faces", the combined image looks like the superposition of multiple transparent scenes. Such linear combinations can not be used for an accurate approximation of natural images. A better approximation can be achieved when the brightness distributions of a larger number (e.g. 50 or 100) images are linearly combined.

Vetter and Poggio [22] have shown that linear combinations of few prototypical example images, that approximate similar images very well, can be obtained when the linear combination is defined by morphing. For this purpose, correspondence is calculated between the prototypical images and a (similar) reference image using a standard optic flow correspondence algorithm. This calculation results in a vector field of displacements that maps the points in the reference image onto the corresponding points of the learned prototypical image. If the feature positions (e.g. of the pixels) in the learned image are given by the vector \mathbf{x} , and if the positions of the same features in the reference image by the vector \mathbf{x}_0 , the result of the calculation of correspondence is the spatial shift vector $\boldsymbol{\xi} = \mathbf{x} - \mathbf{x}_0$ that maps the corresponding feature points onto each other. To define linear combinations of images, Vetter and Poggio proposed to combine the correspondence vector fields

linearly rather than the brightness values of the images¹. If the calculation of the correspondence between the prototypical image p and the reference image yields the spatial shift vector $\boldsymbol{\xi}_p = \mathbf{x}_p - \mathbf{x}_0$, the linear combination of the P prototypical images is defined mathematically by the equation:

$$\boldsymbol{\xi} = \sum_{p=1}^P c_p \boldsymbol{\xi}_p \quad (1)$$

The linearly combined shift vector $\boldsymbol{\xi}$ specifies the shifts of the feature positions in the linearly combined image relative to the reference image. Given the reference image, the linearly combined image is obtained by warping the reference image with the spatial shift vector $\boldsymbol{\xi}$. Such warps of the reference image provide typically very good approximations of natural images. This permits the definition of learning-based representations using only a small number of learned prototypical patterns. The price for the good generalization properties of morphable models is a high computational cost by the necessity to calculate the optic flow.

Given a new image, one can calculate the correspondence of the image with the reference image yielding a shift vector $\boldsymbol{\xi}$. Using equation (1), one can then approximate this shift vector by a linear combination of the prototypical shift vectors $\boldsymbol{\xi}_p$ in order to fit the morphable model to the new image. This permits to characterize the new image by the low-dimensional linear weight vector $\mathbf{c} = [c_1, c_2, \dots, c_P]^T$. On the basis of this vector, the image can be classified. Additionally, nonlinear mappings between the vector \mathbf{c} and interesting parameters that characterize the images, like the pose of faces, can be learned using neural networks [2].

3 Spatio-temporal morphable models

Analogous to the ideas discussed in the last section, morphable models for complex movement patterns or actions should allow to represent classes of similar movements by linear combination of a small number of learned prototypical patterns. Again, we define the linear combination of patterns by the linear combinations of the correspondence fields between the patterns and a reference pattern. In this case, the patterns are not simply stationary images, but image sequences or spatio-temporal trajectories of feature positions. This leads to the theoretical problem to define correspondence between spatio-temporal trajectories. The solution of this problem is the major theoretical contribution of this paper. Spatio-temporal correspondence should be defined in a way that ensures optimal generalization properties of the resulting morphable model.

¹After rewarping the images using the spatial shifts, the brightness values can be additionally linearly combined without causing "transparency effects" (e.g. [22]).

A simple way to define correspondence between image sequences is to calculate usual spatial correspondence between the image pairs that occur at the same points in time. In fact, this strategy has been used in the context of computer animation (e.g. [13, 8, 5]). It leads however not to optimal generalization properties. Natural movement patterns often vary slightly with respect to their timing. It is therefore important to be able to compensate for variations in timing in an efficient way. A simple combination on a frame-by-frame basis does not capture temporal variations in an optimal way, as illustrated schematically in figure 1: Assume two arm movements with the same spatial trajectories, but different timing. At the time t_0 the arm has two different configurations for the two movements that are indicated in white and gray in the figure. If a linear combination of the two movements with equal weights is defined on a frame-by-frame basis an arm configuration arises that is indicated by the dashed line in the figure. The geometrical structure of the arm is strongly distorted, and in particular the upper arm is contracted.

By using more prototypical patterns that are less dissimilar the geometrical distortions could be diminished, but also the storage cost for the prototypes would increase. By defining linear combinations through time warping between the prototypes the necessity to introduce more prototypes can be avoided. In this case, the linear combination at time t_0 then correspond to a natural configuration of the arm (illustrated by the dotted line in the figure) which appears in both sequences at a time that is different from t_0 . All frames of this linearly combined movement correspond to natural configurations of the arm, and the combined pattern has an intermediate timing between the original movements. This motivates a definition of correspondence between movement patterns that also also temporal shifts between the corresponding frames in the two image sequences.

Patterns with different timing

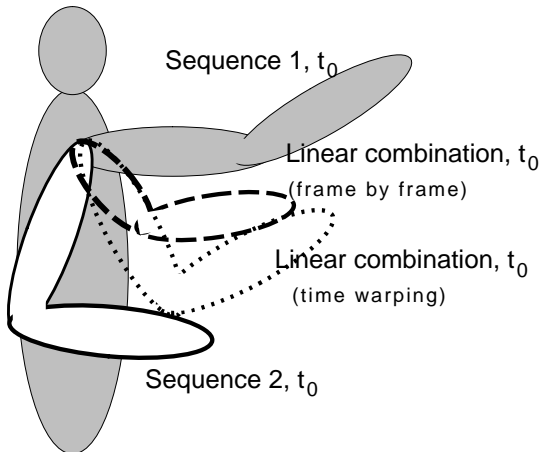


Figure 1: Frame-by-frame correspondence

Mathematically, this can be captured as follows: The first image sequence can be characterized by a trajectory of feature positions $\mathbf{x}_1(t)$ and the second by the trajectory vector $\mathbf{x}_2(t)$. At each point in time t , we define not only spatial shifts $\xi(t)$ (as in the last section) between the corresponding frames, but also temporal shifts $\tau(t)$. These shifts are defined by the following pair of equations:

$$\mathbf{x}_2(t) = \mathbf{x}_1(t') + \xi(t) \quad (2)$$

$$t' = t + \tau(t) \quad (3)$$

The correspondence field consists thus of a spatial and a temporal component which transform the first feature trajectory into the second by spatial warping and time warping.

Having defined what we mean by correspondence between complex movement patterns, we can now construct morphable models analogously to section 2. The linear combination of movement patterns is defined by the linearly combining the spatial and temporal shifts between the prototypical movements and a reference pattern using the same linear weights for spatial and temporal shifts. Let the spatial and temporal shifts of the prototype p be given by $\xi_p(t)$ and $\tau_p(t)$. Then the morphable model is defined by the equations:

$$\xi(t) = \sum_{p=1}^P c_p \xi_p(t) \quad (4)$$

$$\tau(t) = \sum_{p=1}^P c_p \tau_p(t)$$

4 Correspondence Algorithm

Correspondences between two images sequences that are characterized by the feature trajectories $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ are calculated by minimizing the spatial and temporal shifts. Additionally, it is necessary that the calculated correspondences map the patterns in a unique way onto each other according to equation (2). The correspondence algorithm minimizes the following quadratic error functional:

$$E_c[\xi, \tau] = \int [|\xi(t)|^2 + \lambda \tau(t)^2] dt \quad (5)$$

The positive constant λ specifies the trade-off between spatial and temporal shifts and was chosen in order to achieve good interpolation between time-warped sequences, and between sequences with different spatial structure. For the uniqueness of the solution, it must be ensured that the temporal shifts $\tau(t)$ define a smooth one-to-one mapping between the corresponding times t and t' . In addition, for the applications discussed in this paper it will be required

that the first and the last frame of one image sequence are brought into correspondence with the first and the last frame of the other. These conditions lead to the following additional constraints for the optimization problem:

$$d\tau/dt > -1 \quad (6)$$

$$\tau(0) = \tau(t_{\max}) = 0 \quad (7)$$

This constrained optimization problem is most efficiently solved using dynamic programming techniques. The details of the correspondence algorithm exceed the scope of this paper and are described in [10]. The algorithm is very fast and can be easily implemented in real-time when the dimensionality of the feature vector \mathbf{x} is not a extremely large.

5 Classification and estimation of continuous parameters

For the classification and the estimation of continuous parameters that describe properties of movements, first the correspondence between a new pattern and the reference pattern is calculated. This results in a correspondence vector field consisting of the spatial shifts $\xi(t)$ and the temporal shifts $\tau(t)$. Using equation (4) the correspondence vector field is approximated by a linear combination of the correspondence vector fields of the prototypes. The linear weights of the superposition were estimated by minimizing the quadratic error function:

$$E_a(\mathbf{c}) = \int \left| \xi(t) - \sum_{p=1}^P c_p \xi_p(t) \right|^2 dt + \lambda_a \int \left(\tau(t) - \sum_{p=1}^P c_p \tau_p(t) \right)^2 dt \quad (8)$$

$\lambda_a > 0$ specifies the trade-off between temporal and spatial errors for the linear approximation. It was chosen in a way that ensures a good approximation of the spatial and the temporal shifts by the linear combination. The solution of this minimization problem is found by solving a simple linear equation system (for the details see [10]). The estimated linear weights c_p were then used for classification and for the estimation of continuous parameters.

For classification a discriminating function was constructed. Let \mathbf{c}^p be the weight vector that is estimated when the prototypical patterns p is used as test pattern itself. Let \mathbf{c} be the estimated weight vector of a new pattern, and if I_k the index set that contains all indices of prototypes for

the locomotion pattern k (for instance walking or running). Then a discrimination function for the pattern k is given by:

$$D_k(\mathbf{c}) = \frac{\sum_{r \in I_k} \exp(-|\mathbf{c} - \mathbf{c}^r|^2/2\sigma^2)}{\sum_{q=1}^P \exp(-|\mathbf{c} - \mathbf{c}^q|^2/2\sigma^2)} \quad (9)$$

σ is a positive parameter that determines the sensitivity of the discrimination. The quantity D_k is always between zero and one.

To estimate parameters that characterize the movement pattern, like the locomotion direction, a mapping is learned between the linear weight vector \mathbf{c} and the interesting parameter θ . If the mapping is a smooth function it can be approximated by a radial basis function network (e.g. [11]) with the form:

$$\hat{\theta} = f(\mathbf{c}) = \sum_k f_k \exp(-|\mathbf{c} - \mathbf{c}^k|^2/2\sigma_f^2) \quad (10)$$

Such networks were constructed separately for each class of locomotion patterns. The networks were trained with example patterns for which the true parameter values (locomotion direction, strength of marching) were known. By solving a simple linear equation system this leads to the values of the weights f_k of the network. (For details see e.g. [11]). Finally, the estimates of the subnetworks were combined by choosing the estimate of the subnetwork of the locomotion pattern that corresponds to the classification result.

6 Tracking data and preprocessing

We tested the morphable model for movement patterns by trying to classify and to quantify different forms of locomotion. The patterns were recorded with a Kodak DCX-VR1000 camera in a seminar room. Different types of locomotion were performed (walking, running, limping, and marching). Walking and limping were executed in five different directions relative to the camera axis (0, 11.3, 22.5, 33.8 and 45 deg relative to an axis orthogonal to the view direction of the camera. For the 0 deg condition the walking person is seen exactly from the side.) Five different styles of marching were executed that were supposed to cover a continuum with different degrees of "marching-like" appearance, starting with usual "walking", and ending with a kind of Russian "goose-step". For each locomotion pattern multiple repeats were recorded in order to obtain a measure for the reliability of the classification and parameter estimation.

Twelve characteristic points of the figure were "tracked" by hand-marking them with an interactive program. An example video sequence and the tracked feature points are shown in figure 2. The average translatory movement of the walker was estimated by fitting a linear function of time to the hip positions with a least squares fitting procedure.

The estimated translation was then subtracted from the feature positions such that the hip position coincided with the center of the coordinate system. In addition, variations in scaling that resulted from the variable distance of the subject from the camera were compensated by normalizing the distance between hip and head to one. Occluded feature points were added by smooth interpolation of the feature trajectories. To reduce the noise level of the feature trajectories, we normalized the cycle times of the movements to 21 discrete frames, and fitted each individual feature trajectory by second order Fourier series.

For testing the reliability of the method with larger data sets, we generated an additional artificial data set by simulating a three-dimensional stick figure that performed the three locomotion types: walking, running and limping. The time course of the joint angles of the figure was specified by second order Fourier series with parameters that were adjusted such that the movement looked natural. The pattern for limping was created by time warping the pattern for walking using a monotonic distortion function. This pattern was also used to test whether the correspondence algorithm can recover correctly the applied temporal warping function. Using parallel projection, the three-dimensional stick figure was then projected onto a two dimensional image sequence. Different view angles of the camera relative to the figure could be specified for this projection (camera axis rotated to the side between $+25$ deg and -25 deg, and up and down within the same angle regime). In total, over 150 such patterns were generated to evaluate the performance of the method.

For the calculation of the correspondences we used a centered reference pattern that was obtained by first calculating correspondence between all prototypes and one prototype (walking orthogonal to the view axis of the camera). The resulting correspondence fields were then averaged to define a "centroid" of all prototypes, that was used to adjust all spatial and temporal shifts for this centered reference.

We tested that the correspondence algorithm correctly determines shifts that transform the corresponding trajectories into each other by spatio-temporal warping. We also tested that for patterns that are created by time warping the algorithm recovers the applied warping function.

7 Results

7.1 Classification

Figure 3 shows the classification results for 108 simulated test patterns. In this simulation, we used fifteen different prototypes for the locomotion patterns walking (W), running (R), and limping (L) with different view angles of the camera. The figure shows the obtained values for the discrimination functions D_k for walking (solid line), running (dashed line), and limping (dashed-dotted line). With a threshold of value for D_k of 0.5 all test patterns would be classified correctly. The classification seems to be very robust.

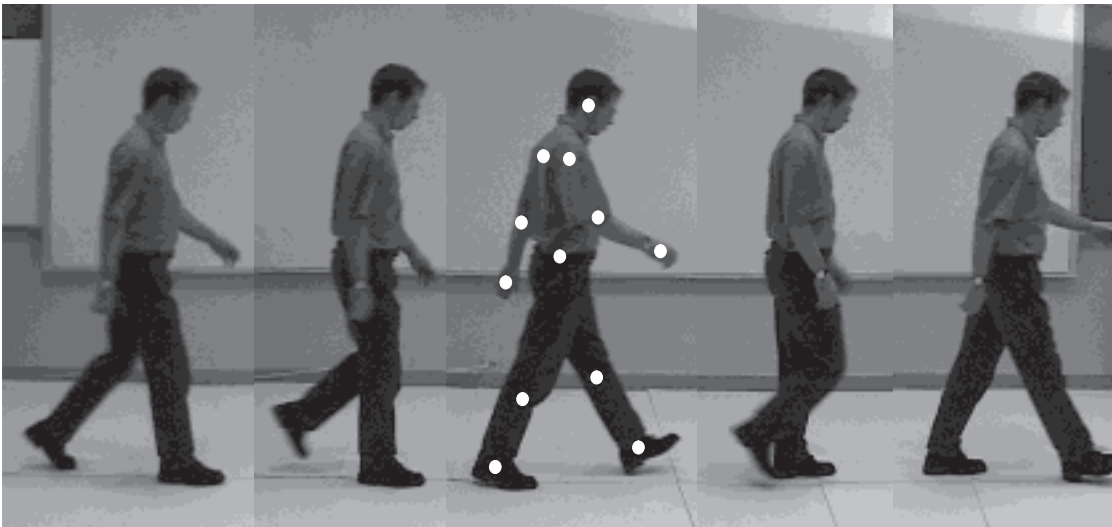


Figure 2: Example image sequence

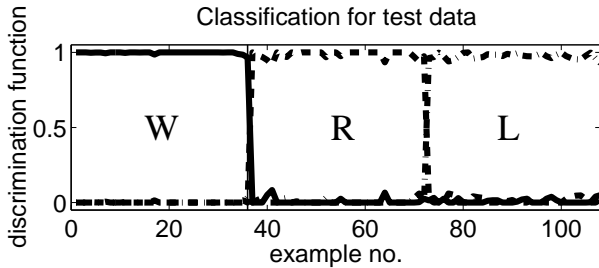


Figure 3: Classification results for synthetic data

The results for real video tracking data are shown in figure 4 for the four locomotion patterns walking (W), running (R), limping (L), and marching (M) for twelve test examples of real video sequences. All test examples were from trials that were not used for the training of the system. A threshold value for D_k of 0.5 (dashed line in the figure) would again lead to correct classification of all patterns. The variability of the discriminating functions is higher in this case than for the simulated data. The discrimination is, however, still very robust.

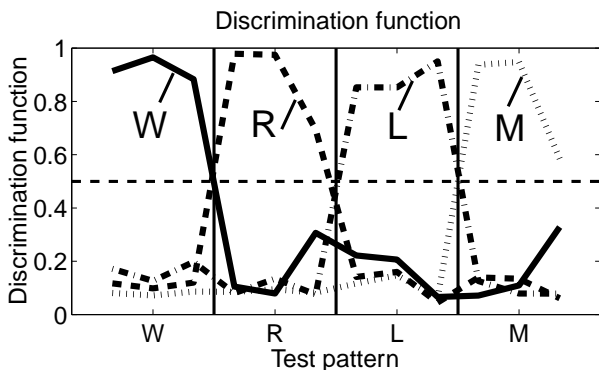


Figure 4: Classification results for real data

7.2 Estimation of continuous parameters

Our method allows not only the classification of movement patterns, but in particular also to estimate continuous parameters that characterize the locomotion. This can be simple geometric parameters, like the locomotion direction relative to the camera, but also more abstract continuous dimensions that describe locomotion patterns. We describe first the estimation of geometric parameters since here an evaluation of the accuracy of the estimation is possible by comparing the estimated with the known true geometrical parameter values.

Figure 5 shows the view direction of the camera relative to the locomotion (side rotation ϕ and top/down rotation θ) for 108 simulated locomotion patterns that all realized different view angles. The system was trained with only fifteen prototypical patterns (walking, running and limping and with the angles $-25, 0, 25$ deg for the rotation to the side, the up-down rotation being zero, and with the same angles for the up-down rotation the side rotation being zero). The black curves indicate the true view directions and the gray curves the estimates obtained from the radial basis function network. The estimates were obtained with fifteen different prototypes with the three locomotion styles walking, running and limping. The estimates are relatively accurate (± 7 deg), which would be sufficient for most surveillance applications.

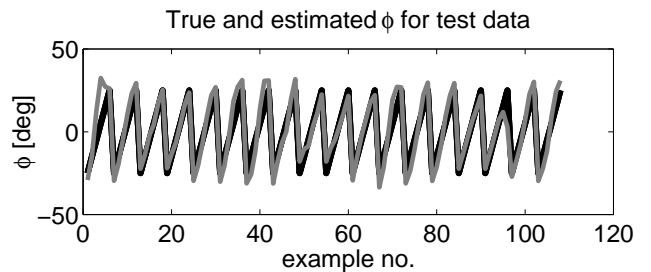
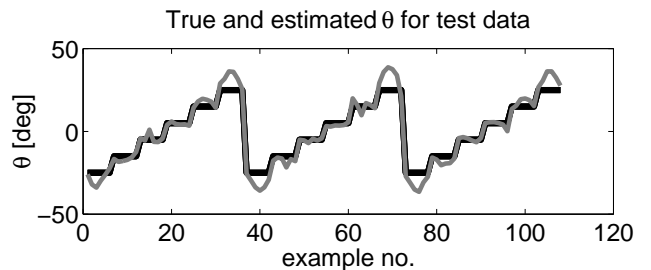


Figure 5: Estimation of the view direction

Figure 6 shows the estimated locomotion direction for "walking" patterns (crosses) and "running" patterns (disks) from real video sequences. Here the network was trained only with the three angles $0, 22.5$ and 45 deg. All other estimates were generated by the generalization properties of the basis function networks. Again all test patterns were new, and none of them was used during the training of the system or as prototype. The variability of these estimated angles is higher than for the simulated data, but the accuracy is still sufficient for obtaining a coarse estimate of the locomotion direction, for instance, in order to detect whether a pedestrian walks toward a street or away from it.

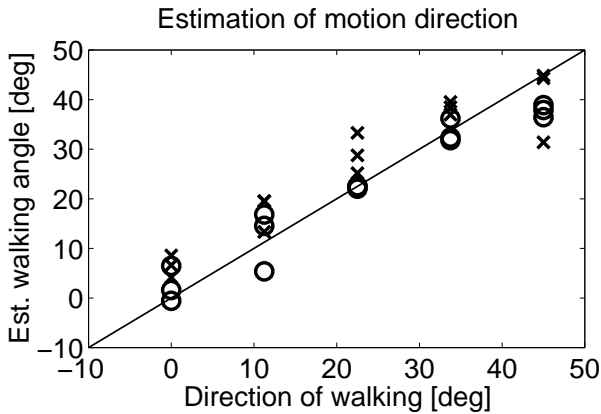


Figure 6: Estimation of walking direction

Another interesting application of our method is the estimation of more abstract parameters that characterize complex movement patterns. As example, we investigated different forms of military marching. By executing different styles of locomotion with increasing "militariness", ranging from regular walking to marching with extensive movements imitating a Russian "goose step", we created a data set with movements that would be assigned different degrees of "militariness" or "intensity of marching" by naive observers. We were interested if our learning-based system would be able to provide a coarse estimate of this abstract dimension: "militariness of marching".

For this purpose we trained the system with walking (W), an example for marching with medium strength (MM), and with an example of Russian "goose step" (SM). These prototypes were associated with the values 0, 0.5 and 1 of an abstract "degree of marching" parameter θ . Figure 7 shows the data from 20 test sequences. During recording, five different degrees of militariness were executed that are ordered along the horizontal axis. Along the vertical axis, the estimated "degree of marching strength" θ is displayed. Remark that the system not only assigns approximately the right parameter values to test patterns that reproduced the locomotion patterns that were used as prototypes and for training of the neural networks (W, MM and SM). It also assigns intermediate values of marching strength to the two other intermediate locomotion styles that were not used during training. Again none of the test patterns was part of the training data. Our method is thus able to recover relatively abstract properties of complex movement pattern, at least if they covary with sufficient reproducible variations in the spatio-temporal structure of the visual patterns.

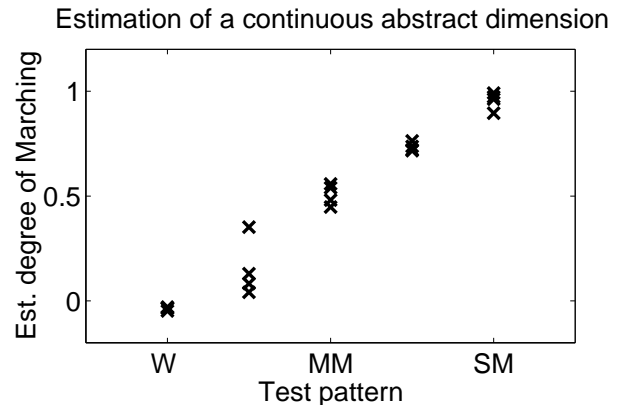


Figure 7: Estimation of the "degree of marching"

8 Conclusions and outlook

In this article we have presented results for a new method for the generation of learning-based representations for complex movement patterns based on a small number of learned prototypical examples. Contrary to many known methods for the recognition of actions, our method allows not only the classification of patterns, but also the estimation of parameters that quantify the spatio-temporal properties of movements along continuous dimensions. Such dimensions can describe simple geometric properties of movements, like the locomotion direction, but also much more abstract properties of movements, like the style of the execution of a certain action.

In the context of surveillance applications, the extraction of continuous parameters seems to have a broad spectrum of applications. The estimation of the locomotion direction could be used to detect if a person is starting to move in the direction of a critical object or a dangerous place (e.g. a street). Information about the specific style of locomotion could be used, for instance, to obtain information about the intention of a walking person (e.g. if it is following another person or planning to attack somebody).

To obtain a system that can be integrated in automatic surveillance systems, major technical improvements still have to be made. A central aim of further work is to link the existing algorithm to a robust method for automatic tracking. Such methods have been proposed in the literature (e.g. [4, 23].) An interesting alternative strategy is motivated by a recent results on the recognition of facial expressions [15]. In this system, the mapping between the outputs of a Gabor filter bank onto low-dimensional weights of a morphable model for stationary images was learned using support vector regression. Using such low-dimensional weights instead of tracked feature position as input for our algorithm would permit to build a system that works in real-time.

For many other technical applications, e.g. in psychophysics, medicine and sports, tracking data from commercial tracking systems are available. The method in the present form can immediately be applied, e.g. to evaluate and to correct complex movements in sports and physiotherapy. In addition, future work is also planned that will investigate if even more subtle information of movements can be extracted. It is known from psychophysical experiments (e.g. [14]) that humans are able to perceive the gender and sometimes even the identity of walkers based on their movements. Compared to other preliminary solutions for this problem (e.g. [17, 16, 12]) our method may not only allow to classify individual according to their movements, but also to extract more subtle psychologically relevant parameters, like whether a person is in hurry, walking in a very relaxed way, or if it has a male or female walking style. We presently test technical system that can extract such subtle information from video tracking data based on the method described in this paper.

Acknowledgments

I thank A. Benali, M. Riesenhuber and C. Nakajima for help with the data acquisition. This work was supported by the Deutsche Forschungsgemeinschaft Gi 305 1/1. Work at CBCL is supported by Office of Naval Research contract No. N00014-93-1-3085 and N00014-95-1-0600, ASI-92-17041. Additional support is provided by: AT&T, Central Research Institute of Electric Power Industry, Eastman Kodak Company, Daimler-Benz AG, Digital Equipment Corporation, Honda R&D Co., Ltd., NEC Fund, Nippon Telegraph & Telephone, and Siemens Corporate Research, Inc.

References

- [1] T. Ahmad, C. J. Taylor, A. Lanitis, and T. F. Cootes. Tracking and recognizing hand gestures, using statistical shape models. *Image and Vision Computing*, 19:(in press), 1997.
- [2] D. Beymer and T. Poggio. Image representations for visual learning. *Science*, 272:1905–1909, 1996.
- [3] M. J. Black and A. D. Jepson. Eigen tracking: robust matching and tracking of articulated objects using a view-based representation. In *Proceedings of the European Conference on Computer Vision, Cambridge*. Springer, NY, 1996.
- [4] A. Blake and M. Isard. *Active Contours*. Springer, New York, 1998.
- [5] V. Blanz and T. Vetter. Morphable model for the synthesis of 3d faces. In *Proceedings of SIGGRAPH 99, Los Angeles*, pages 187–194, 1999.
- [6] J. W. Davis and A. F. Bobick. The representation and recognition of action using temporal templates. Technical Report 402, Massachusetts Institute of Technology, Cambridge, MA, 1996.
- [7] I. A. Essa and A. P. Pentland. Coding, analysis, interpretation and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:757–763, 1997.
- [8] T. Ezzat and T. Poggio. Miketalk: A talking facial display based on morphing visemes. In *Proceedings of the Computer Animation Conference, Philadelphia, PA*, 1998.
- [9] D. M. Gavrila. The visual analysis of human movement: a survey. *Computer Vision and Image Understanding*, 73:82–98, 1999.
- [10] M. A. Giese and T. Poggio. Synthesis and recognition of biological motion patterns based on linear superposition of prototypical motion sequences. In IEEE, editor, *Proceedings of the MVIEW 99 Symposium at CVPR, Fort Collins, CO*, pages 73–80. IEEE Computer Society, Los Alamitos, 1999.
- [11] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural network architectures. *Neural Computation*, 7:219–269, 1995.
- [12] P. S. Huang, C. J. Harris, and M. S. Nixon. Human gait recognition in canonical space using temporal templates. *IEEE Proceedings on Visual Image Signal Processing*, 146:93–100, 1999.
- [13] M. J. Jones. *Multidimensional morphable models: A framework for representing and matching object classes*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1997.
- [14] L. T. Kozlowski and J. E. Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Perception and Psychophysics*, 21:575–580, 1977.
- [15] V. Kumar and T. Poggio. Learning-based approach to real-time tracking analysis of faces. Technical Report 1672, Massachusetts Institute of Technology, Cambridge, MA, 1998.
- [16] J. J. Little and J. E. Boyd. Recognizing people by their gait: the shape of motion. *Videre*, 1:2–32, 1998.
- [17] S. A. Niyogi and E. H. Adelson. Analyzing and recognizing walking figures in XYT. Technical Report 223, Massachusetts Institute of Technology, Cambridge, MA, 1994.
- [18] J. O'Rourke and N. I. Badler. Model-based analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2:522–536, 1982.
- [19] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of The Optical Society of America A*, 4(3):519–524, 1987.
- [20] T. Starner and A. P. Pentland. Recognition of american sign language using hidden markov models. In *International Workshop on Automatic Face and Gesture Recognition*. IEEE, 1995.
- [21] T. Vetter. Synthesis of novel views from a single face image. *International Journal of Computer Vision*, 28(2):103–116, 1998.
- [22] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 19(7):733–742, 1997.
- [23] Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. *Computer Vision and Image Understanding*, (in press), 1999.