Visual Learning and Recognition of a Probabilistic Spatio-Temporal Model of Cyclic Human Locomotion

Miha Peternel and Aleš Leonardis
University of Ljubljana, Faculty of Computer and Information Science
Tržaška 25, 1000 Ljubljana, Slovenia
miha.peternel,ales.leonardis@fri.uni-lj.si

Abstract

We present a novel representation of cyclic human locomotion based on a set of spatio-temporal curves of tracked points on the surface of a person. We start by extracting a set of continuous, phase aligned spatio-temporal curves from trajectories of random points tracked over several cycles of locomotion in a monocular video sequence. We analyze a PCA representation of a set of cyclic curves, pointing out properties of the representation which can be used for spatio-temporal alignment in tracking and recognition tasks. We model the curve distribution density by a mixture of Gaussians using expectation-maximization algorithm. For recognition, we use maximum a posteriori estimate combined with linear data adaptation. We tested the algorithms on CMU MoBo database with favourable results for the recognition of people "by walking" from monocular video sequences captured from the side view.

1. Introduction

Several methods have been developed in the field of computer vision for view based tracking and identification of human locomotion: [9, 6, 2, 5] etc. The approaches can roughly be divided into top-down and bottom-up. The top-down approaches typically assume a spatial model and estimate temporal evolution of configuration parameters. The bottom-up approaches attempt to build a spatial model from a set of primitives and continue with temporal evolution.

Spatial models found in the literature are generally image based or geometry based, either 2D or 3D. Cardboard models put 2D image templates in a geometrical structure. For gait recognition, several authors merely track some predetermined features and do not model the entire view space. There have been several attempts of using image based subspace methods [2] and image statistics [10] to decrease the dimensionality of spatial representation, but these methods tend to lose local description power. Most of the methods are limited in representation of objects that deviate from the assumptions about the spatial configuration.

Temporal evolution is typically represented by a time series of parameters or by state transitions usually modeled by hidden Markov models. State based methods can easily represent a multitude of different motions, but they are not especially suited to capture details of motion, because too many states would be required to model all possible configurations and local variation.

A lot of work on human motion analysis was performed on data accumulated from motion capture of markers attached to human actors. Cedras and Shah [3] discuss experiments by Johansson and others showing that people successfully recognize human motion from a very small set of markers even in the presence of noise.

There have been few attempts of markerless bottom-up structure-free learning of locomotion. Niyogi et al.[8] used spatio-temporal manifold produced by evolution of edges of human silhouette over time, however they do not model motion inside silhouette, neither do they model vertical component of motion explicitly. Torresani et al.[11] attempt to learn moving shape from video.

To the best of our knowledge, there have been no attempts to model human locomotion as a space-time manifold of the whole observable surface. The main advantage of such a model is an ability to probabilistically represent structure and motion in a single framework, with the possibility to include local and global variation.

We assume that the appearance of motion of an articulate object can adequately be represented by a set of trajectories of points on the surface on the object, if the number of tracked points is adequate to approximate the moving surfaces with sufficient degrees of freedom. We focus on cyclic human locomotion, for which a number of databases have been accumulated and natural physical constraints allow us to use a simple point-tracking algorithm and filter out non-optimal trajectories.

We present a novel method for representation of articu-

late cyclic motion based on a set of spatio-temporal trajectories of continuously tracked points on the surface of the observed object. We apply the method to visual learning and recognition of human locomotion. We assume no prior information about the distribution of trajectories. The main advantage of the method is that it probabilistically models the motion over both full view space and time. At this point our method only includes continuously trackable surface points, but in principle the model can include any trackable feature.

The main contribution of this paper is a method for learning and recognition of the spatio-temporal distribution of a set of spatio-temporal curves over a number of iterations of cyclic motion. We generalize spatio-temporal trajectories over iterations using principal component analysis. Finally, we approximate the distribution using a mixture of Gaussians. The recognition is implemented with a combination of maximum a posteriori estimate and linear data adaptation.

The paper is organized as follows: Section 2 outlines extraction and representation of a set of phase aligned spatio-temporal curves, Section 3 introduces PCA decomposition of ST-curves and a Gaussian mixture model for probabilistic presentation for learning of a set of ST-curves, Section 4 describes classification for recognition, Section 5 presents experiments on the CMU MoBo database and the last section contains conclusions and outlines work in progress.

2. Curve extraction

We start by performing random point tracking on the moving object extracted from a monocular video sequence by background subtraction, shadow suppression, and morphological filtering. Each tracked point produces a trajectory which is not necessarily connected and sometimes erroneously jumps among parts of the object.

We define one cycle to be the interval which contains two human steps. We detect cycles in short sequences (around 10 cycles) of locomotion by searching for maxima of autocorrelation of trajectories and voting.

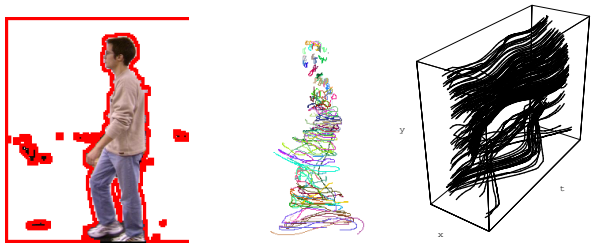

Figure 1. Extraction of a set of phase aligned spatio-temporal curves

We extract only the connected parts of trajectories starting from the beginnings of the detected cycles. The result is a set of phase aligned nearly cyclical spatio-temporal curves (see Fig. 1). We then linearly stitch the curves to make them all cyclic and piece-wise linearly interpolate them to a common size L. We subtract the centroid of the curve from the point coordinates. The intermediate ST-curve representation consists of the curve centroid $\boldsymbol{o} = (o_x, o_y)$ and the centroid-subtracted curve shape vector $\boldsymbol{x} = [x_1, y_1, ..., x_L, y_L]$.

3. Probabilistic spatio-temporal model

In this section we describe learning of a probabilistic spatio-temporal model of a set of ST-curves. The learning procedure is divided in PCA decomposition of the curve shape vectors and subsequent modeling of curve distribution in a subspace by a mixture of Gaussians.

3.1. Curve set in a PCA subspace

Let $\boldsymbol{X}$ be the data matrix with N curve shape vectors $\boldsymbol{x}_n$ of size $D = 2L$ in ordered columns. We perform PCA decomposition according to Anderson [1]:

$$\begin{aligned} \boldsymbol{\mu} &= [\mu_1, ..., \mu_D]^T = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \ , \\ \hat{\boldsymbol{X}} &= \boldsymbol{X} - \boldsymbol{\mu} 1_{1\times N} \ , \\ \boldsymbol{C} &= \frac{1}{N}\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^T \ . \end{aligned} \tag{1}$$

By performing eigenvalue decomposition of the covariance matrix $\boldsymbol{C}$ we diagonalize it $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ in such a way that the orthonormal matrix $\boldsymbol{U}$ contains eigenvectors $[\boldsymbol{u}_1 ... \boldsymbol{u}_D]$ in its columns and diagonal matrix $\boldsymbol{\Lambda}$ contains eigenvalues λ_i on its diagonal, and the eigenvalues and corresponding eigenvectors are arranged in descending order of the eigenvalues. Thus, the most variability of the set of curves is contained in the first few eigenvectors, also called the principal vectors. We use matrix $\boldsymbol{U}$ to remap curve shape vectors $\hat{\boldsymbol{X}}$ on the principal axes: $\boldsymbol{P} = \boldsymbol{U}^T\hat{\boldsymbol{X}}$. The properties of the transform guarantee us that by reducing the representation of the curve shape vector to the first few $d \ll D$ principal components we minimize the reconstruction error in terms of mean square error. The curve representation using a PCA subspace to represent spatio-temporal variation becomes $\boldsymbol{r} = [o_x, o_y, p_1, ..., p_d]$ ($\boldsymbol{o}$ is not transformed).

We analyzed the diagrams of the principal vectors. In all of the cases of a side view the first principal vector is nearly cyclical and contains significant oscillation along the direction of locomotion (see Fig. 2). This feature is further used for phase alignment of curves.

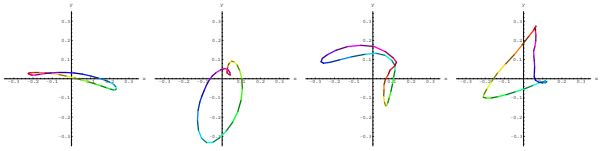

Figure 2. The first 4 principal vectors (left to right)

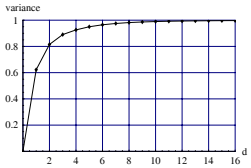

Figure 3. Average variance contained in the first d eigenvectors

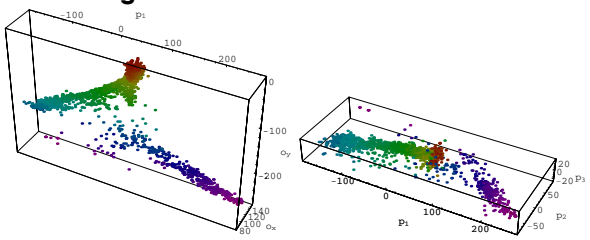

Figure 4. 3D mappings of ST-curve set

3.2. Gaussian mixture model of curve distribution

We approximate the density of the curve distribution with a mixture of Gaussians Θ,

$$\begin{aligned} \mathcal{N}(\boldsymbol{r};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{\frac{|\boldsymbol{r}|}{2}}\sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{r}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{r}-\boldsymbol{\mu})} , \\ p(\boldsymbol{r};\Theta) &= \sum_{i=1}^{K} w_i * \mathcal{N}(\boldsymbol{r};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) . \end{aligned} \quad (2)$$

K is the number of Gaussians and w_i is the weight of Gaussian i with $\sum_{i=1}^{K} w_i = 1$ and $\forall i : w_i \geq 0$. We use diagonal $\boldsymbol{\Sigma}_i$ with $\sigma_1 ... \sigma_K$ on diagonal.

We initialize the means of Gaussians by setting them to a random subset of data vectors. We initialize variances to a random fraction of observed data interval.

We use an iterative expectation-maximization [4] procedure to update GMM:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_j &= \frac{\sum_{i=1}^{N} \boldsymbol{r}_i P(j|\boldsymbol{r}_i,\Theta^s)}{\sum_{i=1}^{N} P(j|\boldsymbol{r}_i,\Theta^s)} , \\ (\hat{\sigma}_j)^2 &= \frac{\sum_{i=1}^{N} (\boldsymbol{r}_i - \hat{\boldsymbol{\mu}}_j)^2 P(j|\boldsymbol{r}_i,\Theta^s)}{\sum_{i=1}^{N} P(j|\boldsymbol{r}_i,\Theta^s)} , \\ \hat{w}_j &= P(j|\Theta) = \frac{1}{N}\sum_{i=1}^{N} P(j|\boldsymbol{r}_i,\Theta^s) . \end{aligned} \quad (3)$$

The final model of observed motion thus consists of principal vectors $[\boldsymbol{u}_1 ... \boldsymbol{u}_d]$ modeling spatio-temporal variation of trajectories and a set of GMM parameters $\{w_i, \boldsymbol{\mu}_i, \sigma_i\}$, $i = 1...K$ modeling distribution density of trajectories in combined 2-dimensional view space and d-dimensional spatio-temporal curve shape subspace.

4. Recognition

The curve extraction steps for recognition are the same as for learning. From that point on we continue with spatio-temporal alignment and maximum a posteriori estimate for classification.

4.1. Spatio-temporal alignment

Given a set of new observation trajectories, we first align them temporally by computing the first principal vector and choose the phase that maximizes correlation with principal vector of the prior model. We account for both original and negated principal vectors, yielding 2 possible results for the phase offset, since we cannot determine the orientation of a principal vector.

We assume that an approximate spatial alignment can be attained by other methods, therefore we can use exhaustive search in a relatively small area of interest, which we perform by linear adaptation of data vectors.

4.2. Classification

We start with the spatio-temporally aligned trajectories from the previous step and recompute a new data matrix of observation vectors $\boldsymbol{Y} = [\boldsymbol{y}_1 ... \boldsymbol{y}_N]$ in the remapped space with principal axes from each prior model c.

We want to find the model c_i that maximizes posterior probability $p(c_i|\boldsymbol{Y})$ given a set of curves $\boldsymbol{Y}$. We use the Bayes rule:

$$\arg\max_i p(c_i|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|c_i)p(c_i)}{p(\boldsymbol{Y})} . \quad (4)$$

Assuming equally likely models c_i and noting that $p(\boldsymbol{Y})$ is the same for all models, the classification simplifies to

$$\arg\max_i p(\boldsymbol{Y}|c_i) . \quad (5)$$

Assuming independence between observations, the final recognition rule is simplified to:

$$\arg\max_i \prod_{n=1}^{N} p(\boldsymbol{y}_n|c_i) = \arg\max_i \sum_{n=1}^{N} \log p(\boldsymbol{y}_n|c_i) . \quad (6)$$

The model c_i that maximizes (6) is selected as the most likely candidate.

Note that computing $p(\boldsymbol{y}_n|c_i)$ requires $\boldsymbol{y}_n$ be remapped to the principal subspace of c_i for recognition, practically

requiring spatio-temporal remapping of data matrix $\boldsymbol{Y}$ and computation of $\sum_{n=1}^{N} \log p(\boldsymbol{y}_n|c_i)$ for each class separately. For the purposes of tracking we may merely be interested in the posterior probability of a single class.

5. Experiments

We used a subset of walk sequences from the CMU MoBo database [7] for experiments. We concentrated on the walk sequences of 25 people captured from the side view, because the trajectories captured from this view exhibit most spatial dynamics potentially useful for recognition. The sequences contain 300-340 frames. Using the described methods we processed the sequences and analyzed their principal vectors (see Fig. 2) and the quantity of variance contained by the first few principal vectors (see Fig. 3). We noticed that the first principal vector contains 38.2%-86.6% of variance, the first 4 vectors contain 84.7%-96.9% of variance and the first 16 vectors contain more than 99.4% of variance. Figure 4 illustrates subspace projections.

We performed curve extraction. The estimated cycle size varied from 28 to 39 frames. We used piece-wise linear interpolation to make all curves of equal length which we defined to be 32 points. We divided the sequences in half, the second half was used for training and the first half was used for testing. We performed PCA and kept 4 principal components to represent spatio-temporal variation. The data vectors thus contained 2 spatial parameters and 4 spatio-temporal parameters. We trained a diagonal GMM with 15 Gaussians using EM algorithm initialized on a random subset of data vectors with random variance scaled from data interval. We used 30 EM iterations on 10 random initializations and kept the best GMM that maximized expectation for each set of the training vectors. We chose the number of Gaussians and iterations empirically.

We tested the classification of 99 test sequences against 99 training sequences using MAP estimate. We tested classification success both within modes and for mixed modes. We phase aligned the sequences by maximizing correlation of the first principal vectors of both sequences. We remapped observation vectors to prior space using only 4 principal components. We tested for different spatial offsets by performing exhaustive search in range $[-32...32]$ for x and $[-16...16]$ for y axis with a step size of 4.

Our method correctly classified most of the sequences as summarized in Table 1. When we inspected the misclassified sequences, we noticed that one misclassified sequence contained deviant arm gestures in the training part of the sequence, while the other misclassified sequence varied in cycle length and included significant positional variations. Mixing modes introduced no additional misclassifications.

Table 1. Summary of recognition results

walk mode	sequences	recognition errors	order of correct class
fast	25	1	2nd
slow	25	0	
incline	25	1	2nd
ball	24	0	
all	99	2	

6. Conclusions

We proposed a novel spatio-temporal model for representation of cyclic human locomotion in monocular view space, together with methods for learning and recognition.

The results for motion based human recognition on a set of 25 people are encouraging. Further tests are required to discount for biases in point tracking and cycle extraction, and to estimate the impact of natural variations in more realistic settings.

The method is currently not scale invariant, but scale can be estimated from the silhouette. In the future we intend to improve the recognition method for scale invariance and varying cycle size.

References

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, 1958.

[2] M. J. Black, Y. Yacoob, and X. S. Ju. Recognizing human motion using parameterized models of optical flow. *Motion-Based Recognition*, pp.245-269, 1997.

[3] C. Cedras and M. Shah. A survey of motion analysis from moving light displays. *IEEE Conf. on Computer Vision and Pattern Rec.*, pp.214-221, 1994.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *JRSSB*, No. 39, pp.1-38, 1977.

[5] J. Deutscher, A. Blake, I. Reid. Articulated Body Motion Capture by Annealed Particle Filtering. *IEEE Int. Conference on Computer Vision and Pattern Recognition*, July 2000.

[6] M. A. Giese and T. Poggio. Quantification and classification of locomotion patterns by spatio-temporal morphable models. *Third IEEE Workshop on Visual Surveillance*, 2000.

[7] R. Gross and J. Shi. The CMU Motion of Body (MoBo) Database. *tech. report CMU-RI-TR-01-18*, Carnegie Mellon University, June, 2001.

[8] S. A. Niyogi and E. H. Adelson. Analyzing gait with spatiotemporal surfaces. In *IEEE Workshop on Nonrigid and Articulated Motion*, pp.64-69, November 1994.

[9] D. Ormoneit, H. Sidenbladh, M. J. Black, T. Hastie. Learning and tracking cyclic human motion. *Advances in Neural Information Processing Systems*, No. 13, pp.894-900, 2001.

[10] P. Saisan, A. Bissacco. Image-based modeling of human gaits with higher-order statistics. *ECCV02*, June 2002.

[11] L. Torresani, A. Hertzmann, C. Bregler. Learning Non-Rigid 3D Shape from Video. *Proc. Of NIPS*, 2003.