

Towards the Automatic Analysis of Complex Human Body Motions

J. Rittscher
GE Global Research
1 Research Circle
Niskayna NY 12308
USA
rittische@crd.ge.com

A. Blake
Microsoft Research
7 JJ Thompson Avenue
Cambridge CB3 0FB
UK
ablake@microsoft.com

S. J. Roberts
University of Oxford
Parks Road
Oxford OX1 PJ3
UK
sjrob@robots.ac.ac.uk

Abstract

The classification of human body motion is an integral component for the automatic interpretation of video sequences. In a first part we present an effective approach that uses mixed discrete/continuous states to couple perception with classification. A spline contour is used to track the outline of the person. We show that for a quasi-periodic human body motion, an autoregressive process is a suitable model for the contour dynamics. A collection of autoregressive processes can then be used as a dynamical model for mixed state Condensation filtering, switching automatically between different motion classes. Subsequently this method is applied to automatically segment sequences which contain different motions into subsequences, which contain only one type of motion.

Tracking the contour of moving people is however difficult. This is why we propose to classify the type of motion directly from the spatio-temporal features of the image sequence. Representing the image data as a spatio-temporal or XYT cube and taking the 'epipolar slices' [4] of the cube reveals that different motions, such as running and walking, have characteristic patterns. A new method, which effectively compresses these motion patterns into a low-dimensional feature vector is introduced. The convincing performance of this new feature extraction method is demonstrated for both the classification and automatic segmentation of video sequences for a diverse set of motions.

1 Introduction

The automatic interpretation of video sequences is relevant to a growing number of applications of digital video technology. The automatic annotation of image sequences for example is very important for archiving film and video. The interpretation of human actions plays a vital role for intelligent environments and surveillance applications. In the case of the surveillance application this classification has to be instantaneous such that an alarm message can be generated once a suspicious action is observed. Due the high dimensionality of the configuration space, the tracking of people with a fully three-dimensional articulated motion model [7, 23] from a single view is difficult. Other than in the case of human motion capture the aim of this work is to classify the type of motion without having to estimate the pose. Therefore we avoid the use of an articulated model. The guiding principle of the work presented here is to use a low dimensional representation for modelling appearance and to facilitate the instantaneous classification of the observed motion. Motion classification experiments based on two different approaches are presented. The first approach uses the technique of active contour tracking to simultaneously perceive and

classify the type of motion. Later a method which classifies the type of motion directly from the set of spatio temporal features of the sequence is presented.

Modelling people by using their apparent contour as used by Baumberg and Hogg [1] is a compromise between a complex articulated model and a basic blob tracker. In the contour tracking framework [3] the apparent contour of an object is modelled as a spline contour. The deformation of this contour with respect to a template is controlled by a low dimensional linear state space, the shape space. The dynamics of the contour is then modelled by a stochastic process. A number of researchers [24, 5, 27] have investigated the use of Hidden Markov Models (HMM) to describe complex motions. Wilson and Bobick [28] as well as Starner and Pentland [24] use a Hidden Markov model to recognise hand gestures. Here the observation probabilities of the HMM model pixel coordinates. The recognition of gestures is different from recognising motions as hand gesture do not depend generally on the speed of the motion. But speed or frequency of a motion provide a strong clue for detecting the class of motion. Bregler [5] presents a system which tracks and classifies different motions. Four different levels of abstraction are established each supporting a set of hypotheses. On the highest level a number of Hidden Markov models are used to evaluate which complex gesture is observed. The states of the HMM correspond to different dynamical motion models. But the models are only used to describe very simple and elementary motions like moving left and right, up or down. It will be demonstrated here that statistical models used for the anticipation of motion (linear, stochastic differential equations and their discrete embodiment as auto-regressive processes) are capable of modelling certain repetitive human motions such as running and walking.

In most approaches tracking and motion classification are dealt with as two separate processes. However it is highly desirable to develop systems where classification feeds back into the perception of motion since perception and classification are inextricably bound together. The reasoning behind the approach is that the statistical models used for the anticipation of motion can potentially be adapted for classification. This concept can be implemented by extending the CONDENSATION [12] filter to a mixed state tracker which automatically switches between different motion models. This is possible since the CONDENSATION filter can handle several different hypotheses at the same time and allows for a nonlinear model to predict the position of an object in the next time step. The current position and deformation of the contour relative to a template is specified through a continuous variable. A discrete variable denotes the current motion model being used. This work is therefore related to [2]. The main difference is that Black and Jepson model motion as a trajectory through a configuration space spanned by a basis of optical flow fields. The computation of the optic flow field for the entire area of interest requires considerable computational complexity.

An important aspect of this work is that by switching automatically between different motion classes an automatic segmentation of the sequence is obtained. The duration of the different motions are explicitly modelled by a Markov chain which applies long-term temporal constraints. Because of the expected length of a duration of one particular motion the probabilities of state transitions tend to be very small. As a consequence only a small number of particles will be propagated with an alternative motion model. In order to avoid the use of an extremely high number of samples a *Partial Importance Sampling* (see section 3.2) strategy is developed to approximate the posterior probability with a low number of samples. One important fact is that the importance sampling method used here has $O(N)$ complexity and not $O(N^2)$ as in [13]. It will be demonstrated that this framework can also be used to learn complex dynamics from a sequence containing mixed motions.

2. Simultaneous perception and classification

When tracking with an active contour [3] the objects apparent contour is described by a spline curve. The time-varying shape of the spline contour with respect to a template at time t is specified in terms of a continuous state vector $x_t \in \mathbb{R}^{N_x}$. So the problem is to estimate the state vector x_t based on a set of measurements z_t taken from the current frame of the image sequence. But rather than having a single estimate in the Bayesian framework the posterior probability $p(x_t|Z_t)$ is propagated over time. Z_t denotes the history of the measurements, i.e $Z_t = \{z_0, \dots, z_t\}$. Here the probability density $p(x_t|Z_t)$ is represented in a non-parametric form by a set of K samples $\{x_t^{(i)}\}_i$. Each sample $x_t^{(i)}$ has a likelihood weight π_i associated to it, such that $\sum_i \pi_i = 1$. The interpretation of such a particle set is that if the set is *resampled*, meaning that an X is chosen to be one of $X_t^{(n)}$, with probability proportional to its weight $\pi_t^{(n)}$, that X is distributed (approximately) according to the posterior p_t . A Condensation or particle filter [10] is applied to propagate $p(X_t|Z_t)$ over time. Multi-class dynamics are represented by appending to the continuous state vector x_t , a discrete state component y_t to make a ‘‘mixed’’ state $X_t = (x_t, y_t)^T$, where $y_t \in \{1, \dots, Y\}$ is the discrete component of the state, labelling the class of motion. Corresponding to each state $y_t = y$ there is a dynamical model, taken to be a Markov model of order K that specifies $p_i(x_t|x_{t-1}, \dots, x_{t-K})$. A linear-Gaussian Markov model of order K is an autoregressive Process (ARP) [16] defined by

$$x_t = \sum_{k=1}^K A_k x_{t-k} + \mathbf{d} + B\mathbf{w}_t \quad (1)$$

in which each \mathbf{w}_t is a vector of N_x independent random $N(0, 1)$ variables and $\mathbf{w}_t, \mathbf{w}_{t'}$ are independent for $t \neq t'$. Each class y has a set (A^y, B^y, \mathbf{d}^y) of dynamical parameters, and the goal is to learn these from example trajectories. Note that the stochastic parameter B^y is a first-class part of a dynamical model, representing the degree and the shape of uncertainty in motion, allowing the representation of an entire distribution of possible motions for each state y . In addition, and independently, state transitions are governed by the transition matrix for a 1st order Markov chain:

$$P(y_t = y' | y_{t-1} = y) = M(y, y').$$

Observations \mathbf{z}_t are assumed to be conditioned purely on the continuous part x of the mixed state, independent of y_t , and this maintains a healthy separation between the modelling of dynamics and of observations. Observations are also assumed to be independent, both mutually and with respect to the dynamical process. The observation process is defined by specifying, at each time t , the conditional density $p(\mathbf{z}_t|x_t)$ which is taken to be Gaussian in experiments here. The Condensation filter, or particle filters [8, 15, 11] in general, can be extended straightforwardly [12] to deal with mixed states. A maximum likelihood learning approach as presented in [19] is used to estimate the set of dynamical parameters (A^y, B^y, \mathbf{d}^y) .

3 Automatic segmentations

Autoregressive processes are a special class of Gaussian Markov processes. It is known that a second order Markov process in continuous time is governed by a stochastic differential equation, the Fokker-Plank or Kolmogorov equation [14]

$$m\ddot{x} + c\dot{x} + kx = bw ,$$

where w is white noise. When there is no noise present, i.e. $w = 0$ this is an oscillator with mass m , damping constant c and stiffness k . The sample path of such a process in continuous time in one dimension is a damped harmonic oscillation. A multivariate process can be decomposed into damped harmonic oscillations along different directions of the configuration

space. A comparison of three different learnt models for walking presented in figure 1 indicates that most of the $2N_x$ degrees of freedom of the second order autoregressive model are 'unused' - i.e. discarded by the learning algorithm.

3.1 Motion models for classification

The continuous variable X_t describes the current position and deformation of the contour relative to a template. In the case where two different motions are already characterised by having very different configurations X_t the classification problem is not very hard because essentially the shape information is used to discriminate between the different motions. Motions which can only be discriminated in phase space make the classification problem difficult. In this case only the dynamical information can be used to discriminate between the two different motions. The motion models need to be finely tuned in order to allow the mixed state CONDENSATION filter to automatically select the correct motion class. Since there can still be ambiguities between different motion models we cannot expect that the dynamical information alone will result in good segmentation results. Therefore as important as tuning the models is to model the expected duration of each motion correctly. The Markov chain with the transition matrix M does in fact act as a prior to the perception of the motion. It will be demonstrated later that an inappropriate set of transition probabilities will result in a very bad segmentation of the sequence.

It is a standard result of Markov chains, that the mean duration of a motion from class y is given by $1/(1 - M(y, y))$. For example, for a human body motion of mean duration 2 seconds, at a video-field rate of 50 Hz, $M(y, y) = 0.99$. Since also, for any Markov transition matrix, $\sum_{y'} M(y, y') = 1$, and all elements positive, we must have $M(y, y') \leq 0.01$ for any $y' \neq y$.

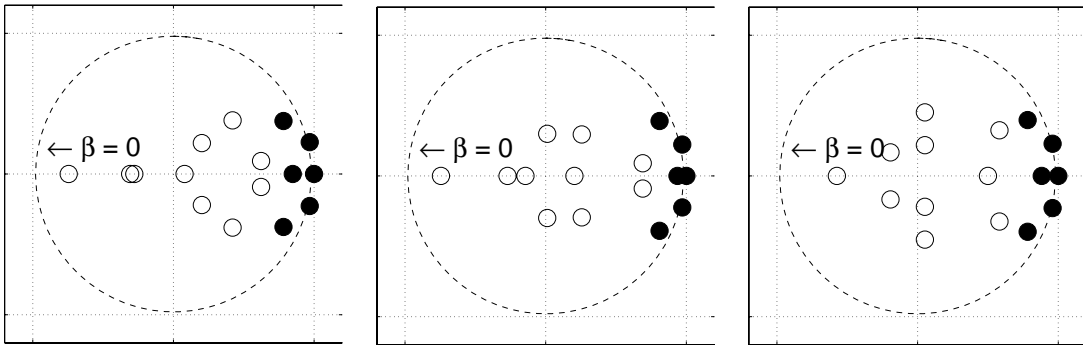


Figure 1: **Discrete eigenvalues of motion models learnt from three different walks.** Three different walking sequences of one human are tracked. For each sequence a second order autoregressive process are learnt using the Maximum Likelihood learning rule. Only the periods in which the walker is in a steady state are used to learn the model. It is crucial to observe that only the first 4 modes (displayed as black circles) are similar for all three models. From these only the first 2 nodes are significant since they have a time constant $\beta^{-1} > 1$ second. All remaining modes have a damping constant $\beta^{-1} < .2$ seconds.

Thus, in a filter using N particles, only $0.01N$ particles are committed at each time step, on average, to the possibility of a change of motion class. Thus it is to be expected that mixed-class tracking could require, in the worst case, that N be hundred times greater than for single-class, for comparable performance. One general approach to reduce the number of particles is to use importance sampling [9], in which areas of configuration space that are unduly sparsely populated with particles can be artificially repopulated, and corresponding likelihood weights $\pi_t^{(n)}$ reduced to maintain the correct posterior distribution. This is done using an importance function $g(X)$ which determines the intensity of repopulation over the configuration space for X . In the standard approach, given a prior $p_0(X)$, a particle set $\{(X^{(n)}, \pi^{(n)}), n = 1, \dots, N\}$, is constructed, in which

the $X^{(n)}$ are drawn from p_0 and likelihood weights are $\pi^{(n)} = p_0(X^{(n)})/g(X^{(n)})$. Then, a random variable X generated by resampling from the particle set is distributed (approximately) as $p_0(X)$.

3.2 Partial importance sampling

In the multi-class classification problem, it is the discrete component y of the state X for which importance sampling is required. Importance sampling of the continuous component x certainly has applications [13], but where it is not required, the sampling algorithm can be simplified as follows. The importance sampling function g now takes the form

$$g_t(X_t|X_{t-1}) = p(x_t|x_{t-1}, y_t)P(y_t|y_{t-1}) ,$$

where $P(y_t|y_{t-1}) = G(y_{t-1}, y_t)$. It mimics the true process dynamics with respect to the continuous x and the discrete component y is sampled according to the importance transition matrix G . Now the forward algorithm, with Partial importance sampling over the discrete component of state only, is as follows.

Algorithm:

1. Choose an index m randomly (with replacement) from $m = 1, \dots, N$, with probability proportional to $\pi_{t-1}^{(m)}$.
2. Choose $y_t^{(n)}$ with probability $G(y_{t-1}^{(m)}, y_t^{(n)})$ from $y = 1, \dots, Y$.
3. Choose $x_t^{(n)}$ by sampling at random from the distribution $p(x_t|x_{t-1}^{(m)}, y_t^{(n)})$, i.e. compute $x_t^{(n)}$ for the AR process with the parameters (A^y, B^y, d^y) where $y = y_t^{(n)}$ that is, for an autoregressive process with order $p = 1$ for the dynamics of class $y = y_t^{(n)}$:

$$X_t^{(n)} = A_1^y X_{t-1}^{(m)} + \mathbf{d}^y + B^y \mathbf{w}_t^{(m)}$$

where the $\mathbf{w}_t^{(m)}$ are vectors of normal random variables generated independently for each m, t .

4. Set $\pi_t^{(n)} = p(z_t|x_t^{(n)}) \frac{M(y_{t-1}^{(m)}, y_t^{(n)})}{G(y_{t-1}^{(m)}, y_t^{(n)})}$

Two different motion classes were used for a classification experiment. A *pure jump*, i.e jumping up and down without lateral arm or leg movement and a *half star* which is a 'star jump' without arm movement. In order to illustrate the two motions the contours of the previous time frames are superimposed on one frame (figure 2). A separate set of training sequences which only contain one type of motion were used to learn the motion models. In all the experiments the tracking process is initialised by hand. In order to get a very good model the Maximum Likelihood learning rule was used on a sequence of 8 seconds length. Note that a combination of these two motions illustrates both types of classification problems: The lateral leg movement makes it easy to decide when the person is in motion class *half star*. The jumping up from the bending down position is on the other hand very difficult. As motivated before an importance sampling method on the discrete state has two advantages. A lower number of particles can be used as well as modelling a low probability of a change of motion class. The importance weights were chosen such that the off-diagonal weights were 0.01 or 0.05, i.e. the probability that a particle changes state is 10 times more likely than without partial importance sampling. As can be seen in figure 4 it was indeed possible to use a lower number of samples to archive a asymptotical misclassification performance of 10%. By using 400 samples it was possible to obtain a very good segmentation of the sequence. The asymptotic error rate however was not reduced by this method. The misclassification for the set of three different test sequences was 10%.

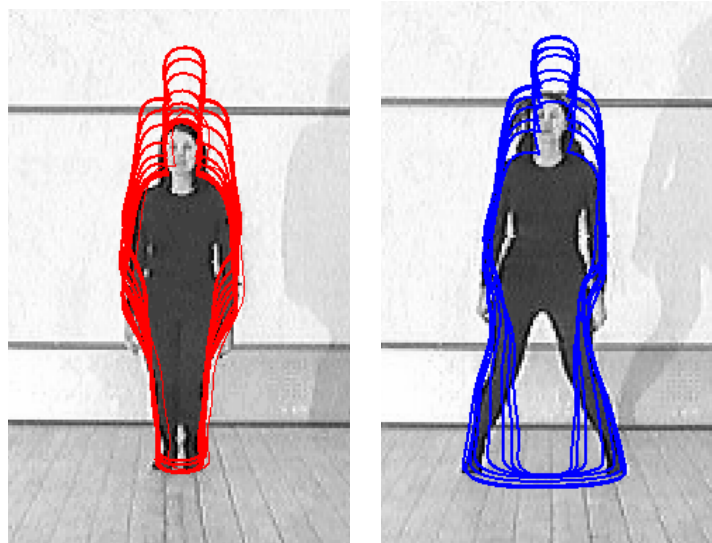


Figure 2: **Motion classes used in the experiment.** In order to illustrate the motion contours of previous time steps are superimposed on one frame. Left image: pure jump, i.e. jumping up and down without lateral arm or leg movement. Right image: half star - a 'star jump' without arm moment. Notice that both motions begin with an upwards acceleration. Hence it is then difficult to discriminate between both motions.

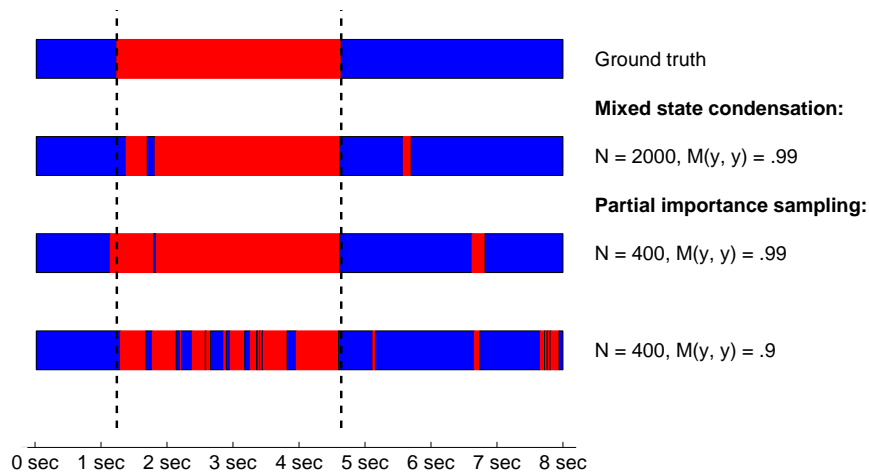


Figure 3: **Examples of the different segmentations.** The motion class pure jump is shown in red, half star in blue. The top row corresponds to the ground truth obtained by hand segmenting the sequence. A two state CONDENSATION filter without importance sampling is used with $N = 2000$ particles. The third row displays the segmentation obtained using partial importance sampling on the discrete state with $N = 400$ samples. Note that the quality of the segmentations does not get worse when the sample size is reduced. The crucial role of the transition matrix is documented in the bottom row. Here a transition matrix with $M(y, y) = .9$ and importance weights with $g(y, y) = .8$ are used, hence the expected duration of the motion is no longer correct.

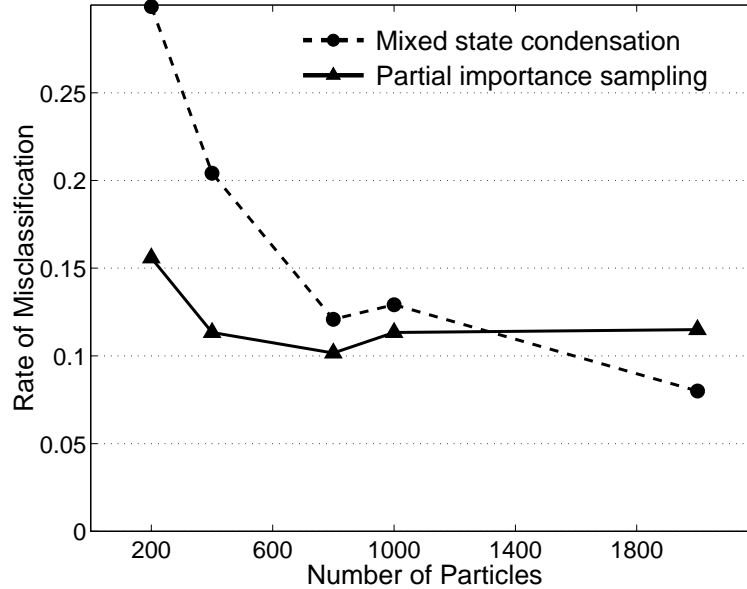


Figure 4: **Misclassification Graphs.** The average error of three tests for each of the three filters on a particular test sequence is displayed for different particle sizes N .

4 Classifying motions directly from spatio-temporal features

The robustness of the tracking process obviously depends on the foreground model. Sullivan [25, 26] *et al.* show that the quality of the tracking can be improved significantly by taking the structure and variability of the foreground into account. It is however difficult to construct such a model of a human body in motion. Yacoob and Black [29] track different body parts and use a vector of measurements over the temporal axis to characterise the type of motion. In order to compensate for a certain variability of the motion they formulate an algorithm which recognises an observed activity subject to a set of admissible transformations. Although this is an alternative which reduces the complexity of the foreground model the problem of tracker initialisation is still open. The main aim of the work presented here is to use a very basic approach to track the foreground region and subsequently recognise the type of motion based on the spatio-temporal features of the image sequence.

One possible spatio temporal feature is optic flow which is used in [2] to recognise gestures and expressions. But because the flow field of the entire image is used these measurements are not very well localised and can easily be corrupted by background motion. Chomat *et al.* [6] present a system which recognises certain motions by evaluating the responses of spatio-temporal filters. Their system is, for example, able to detect if a person enters or leaves a room. We are, however, interested in a much finer distinction of motion patterns. In our experience it is very difficult to discriminate between motions, such as different walking styles, on the basis of spatio-temporal filters. Representing the image sequence as a space time cube [4] as shown in figure 7 reveals that different motions such as running and walking have characteristic patterns. Niyogi and Adelson [17, 18] use this observation to recognise people by their gait. They fit spatio-temporal surfaces to the person in motion in order to estimate the parameters of a stick-figure model. Recognising the type of motion should not require any articulated motion model, or in fact, any intermediate representation. Section 4 of this article presents the construction of

a feature extraction method which projects the data from the three dimensional cube onto a low dimensional space. It will then be shown that this method can be used to classify certain motion patterns. Finally it will be demonstrated that the same method can be applied to classify sequences of mixed motions instantaneously.

In addition to the dynamical model the contour tracking approach requires a suitable observation likelihood and a tracker initialisation. In the case of tracking people both of these problems are difficult to solve. Therefore we now present, as mentioned in the introduction, a technique which allows to recognise a particular motion pattern directly from the spatio-temporal features in the image sequence. The key idea here is to use slices of the spatio-temporal cube, epipolar slices to be exact, to classify the type of motion. The braided pattern of the epipolar slices, as seen in figure 6 clearly contains enough information to discriminate the motion pattern generated by different walking styles. We now aim to design a specific feature extraction method which captures the salient information of this pattern. Due to the regularity of the motion pattern it will be possible to encode the braided pattern using only a small number of parameters.

Motions like running, walking, and skipping can be characterised by the different intrinsic velocities of leg movement. As discussed before, epipolar slices of the spatio-temporal cube (see figure 7) exhibit a braided pattern which characterises the type of motion. Since the epipolar slice is an entity in space-time, the braided pattern is directly related to the velocity-profile of the motion. The braided pattern of the leg motion, for example, consists of two self intersecting curves, one for each leg. Hence the velocity of the leg can be computed by estimating the outer normal to the curve. This is also illustrated in figure 5. The angle α of the outer normal can be computed as

$$\tan(\alpha) = \frac{\delta_t I}{\delta_x I}$$

Since the tan-function is antisymmetric, i.e. $-\tan^{-1}(-x) = \tan^{-1}(x)$, for every time t the modulus of α , $|\alpha|$, estimated from either curve, is identical. Hence the distribution of $|\alpha|$ will describe the velocity distribution of the motion pattern. In

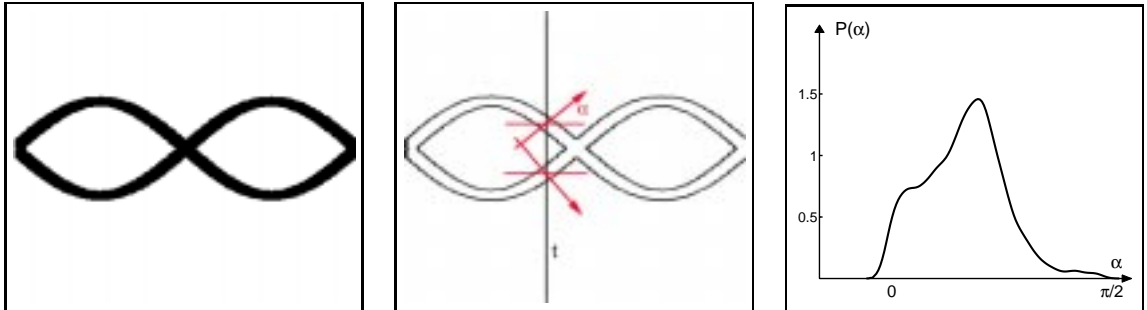


Figure 5: **Idealized braided pattern.** The figure on the left shows an idealised braided pattern of a person walking under a fronto parallel view. The angle of the outer normal to one curve, α , measures the velocity of the leg at time t . The second curve is of course a mirror image of the first hence the angles of both other normals are related. This is because tan-function is antisymmetric, i.e. $-\tan^{-1}(-x) = \tan^{-1}(x)$. It is also clear that the angles of the inner normals correspond to those of the outer normals. The distribution of $|\alpha|$ hence characterises the braided pattern. The graph on the right shows a typical learnt distribution of $|\alpha|$ from an epipolar slice taken from the experimental data.

order to characterise a particular motion it will be necessary to take a collection of epipolar slices into account. It would be therefore desirable to use a more compact representation of the distributions of $|\alpha|$. The learnt distributions, $p(|\alpha|)$, shown in figure 8, have very characteristic shapes. The shape of the distributions depends on both the location of the epipolar slice and the type of motion. They are clearly different in their third moment. The estimate of the third moment or skewness, γ_1 (see

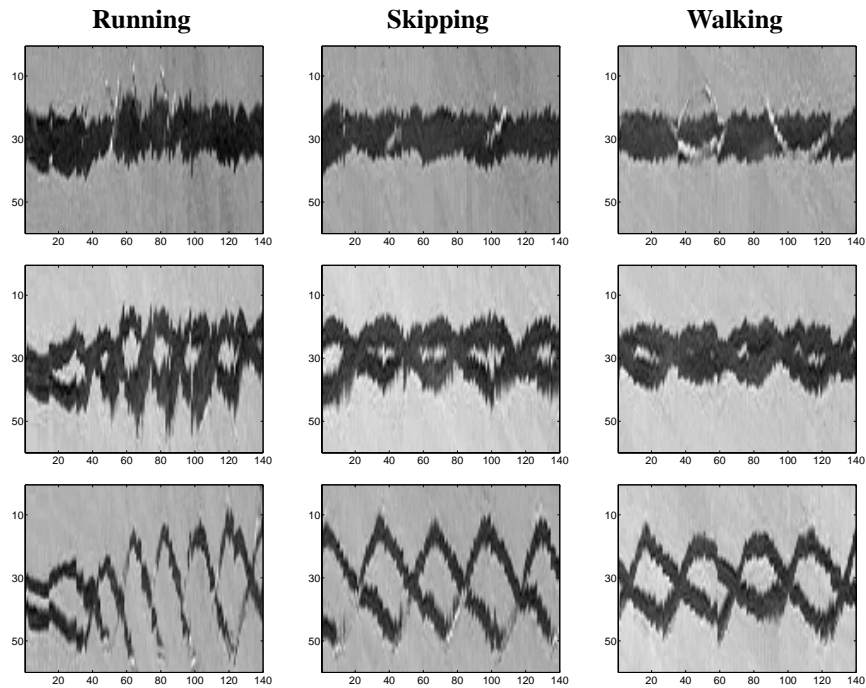


Figure 6: **Epipolar slices of the space time cube.** As indicated in figure 7, this graph displays three typical epipolar slices for the three different motions running, skipping and walking. Note that these epipolar slices are taken from sequences where the centroid of the foreground region, i.e. the walking person, is stationary. The purpose of this preprocessing is to factor out the general motion and isolate the pattern which characterises a particular walking style. The image heights are chosen such that the slices on the top row correspond to hip motion, the middle row to upper leg and the bottom row to the shin. It can be clearly seen that the upper leg and shin move at different speeds.

for example, [22]) of the distributions is defined as,

$$\gamma_1(\{\alpha_1, \dots, \alpha_N\}) = \frac{1}{N} \sum_{j=1}^N \left(\frac{\alpha_j - \bar{\alpha}}{\sigma} \right)^3. \quad (2)$$

The skewness of a distribution measures the degree of asymmetry. The measure does not depend on the location or scale (measured respectively by the mean $\bar{\alpha}$ and the variance σ). Hence a linear transformation of the distribution will not affect the skew factor γ_1 . For a symmetrical distribution the skewness, γ_1 , is evidently zero. A positive value of skewness signifies a distribution with an asymmetric tail extending to the right of the mean and vice versa. Three typical epipolar slices, ranging from fast to slow motions, are displayed in figure 8. Both the learnt distributions and the skewness, γ_1 , allow to discriminate between the three different velocity profiles. We therefore conclude that it is sufficient to compute the skewness of the learnt distribution of a collection of epipolar slices and treat the vector of the skew factors as a feature vector.

4.1 Practical computations

Two steps are necessary to estimate α from real data. Firstly, the motion boundary needs to be extracted and, secondly, α needs to be estimated from noisy data. Based on the assumption that a suitable bandpass filter will be sufficient for denoising, the space time cube is first convolved with a spatio-temporal Gaussian filter, φ . In the next step the partial derivatives $\delta_x I$ and $\delta_t I$ are computed from the bandpass-filtered data. This can, of course, be performed in one step by convolving the data with the partial derivatives of the filter φ . In order to estimate the normals to the self-intersecting curves, α is calculated where $|\nabla I|$ exceeds a threshold, i.e. $|\nabla I(x, t)| > C$, thence

$$\tan(\alpha(x, t)) = \frac{(I * \delta_t \varphi)(x, t)}{(I * \delta_x \varphi)(x, t)}. \quad (3)$$

Unless otherwise stated the threshold C is set naively. In fact the threshold C can be set conservatively. Its only purpose is to prevent the distribution $p(|\alpha|)$ from being swamped by α -values which correspond to locations (x, t) at which the modulus of the gradient $|\nabla I(x, t)|$ is near zero.

Figure 8 displays a set of epipolar slices for running, walking, and skipping. These slices were taken at the height of the shin. Figure 8 also shows the distributions $p(|\alpha|)$ for each of the epipolar slices and the corresponding skew values. One can observe that both, the shape of the distribution $p(|\alpha|)$, and the skew values γ_1 discriminate between the different motions. The distribution $p(|\alpha|)$ clearly depends on the position of the epipolar slice. Hence we conclude that a collection of epipolar slices taken from different heights of the space time cube is necessary to discriminate between the different motions.

In order to formalise the approach it is necessary to make the notation more precise. Since each epipolar slice of the sequence I depends on the height y (see figure 7) we denote the skew factor of the corresponding $|\alpha|$ -distribution as $\gamma_1(I, y)$ and define

$$\gamma_1(I) := \{\gamma_1(I, y_1), \dots, \gamma_1(I, y_N)\}, \quad (4)$$

where N is the number of epipolar slices which is taken from the space time cube. Hence we compress the image sequence I to a feature vector of length N .

So far we have discussed the problem of classifying sequences which only contain one type of motion. One requirement for classifying the motion instantaneously is the estimation of the moments of $p(|\alpha|)$ from a very short time window. Consequently this problem has two different aspects. Firstly, the estimation of the moments from a small number of samples and

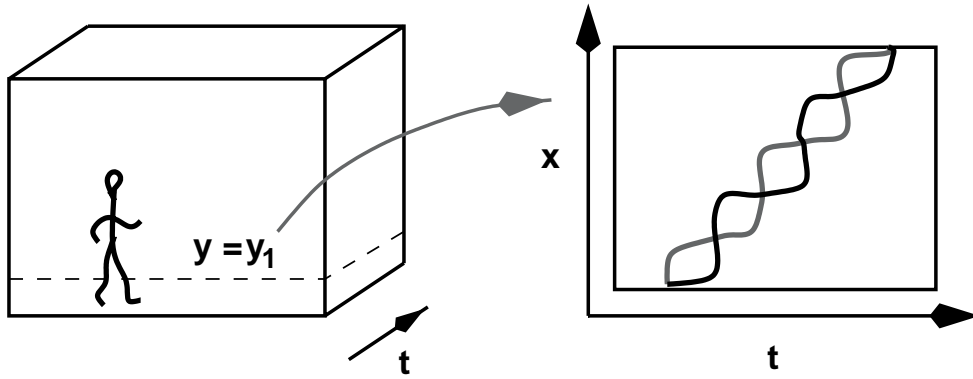


Figure 7: **Spatio-temporal cube or XYT cube.** In order to convolve the image sequence with a spatio-temporal filter the different frames of the sequence are arranged in a so called spatio-temporal cube (left). The width and height of the cube are determined by the width and height of the image. The depth of the cube is determined by the temporal length of the filter. An *epipolar slice* of the XYT cube is defined for a fixed image height $y = y_1$. The effect of taking an epipolar slice of the spatio-temporal cube is illustrated by a person walking across the scene. The leg motion gives rise to a particular braided pattern which is visible in the epipolar slice (right).

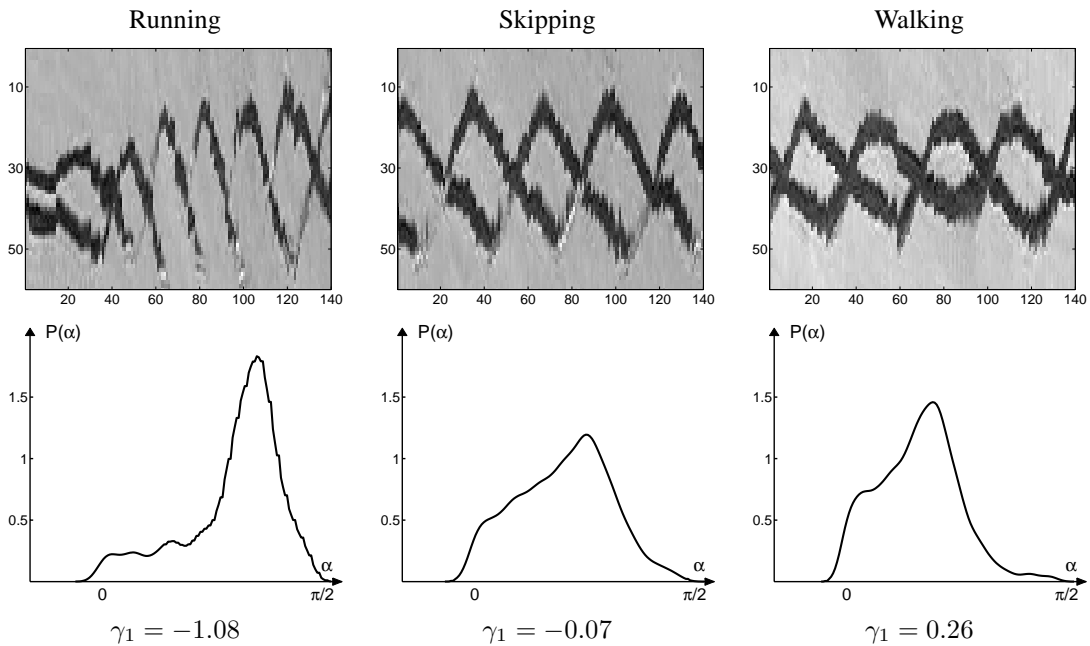


Figure 8: **Skewness of the $|\alpha|$ -distributions.** This figure illustrates the effect of estimating the skewness γ_1 (2) of the learnt distributions for $|\alpha|$. The top row shows examples of typical epipolar slices for running, walking and skipping. The corresponding distributions of $|\alpha|$ are shown in the bottom. It should be noted that all three representations of the data, the raw data, the learnt distribution of $|\alpha|$ and the skewness of the distribution allow us to discriminate the three different motion patterns. In these examples the foreground pattern (in black) is clearly visible. But since this method is based on the estimation of normals the representation does not depend on the intensity difference between foreground and background as long as exceeds the threshold C (see equation (3)).

secondly, depending on the temporal length of the sequences, the dependency on the phase of the motion. In order to obtain more robust estimates on small sample sizes one could try to approximate the distribution $p(\alpha)$ a mixture of Gaussians, where

$$p(x) = \sum_{j=1}^M p(x|j)P(j) , \quad (5)$$

where each of the component densities $p(x|j)$ is chosen to be a normal distribution, i.e. $p(x|j) = N(\mu_j, \sigma_j)$. The coefficients $P(j)$ are referred to as the mixing parameters which sum to unity. The mixing parameters are hidden variables of the model. The mixture of Gaussians allows for the computation of moments in closed form. But our experiments have shown that this method only provides a marginal improvement.

As opposed to the previous method described in section 3 the segmentation boundaries detected by this method depend on the length of the time window for computing the skew vector $\gamma_1(I)$. In case this time-window becomes too short the variance of the learnt distributions $p(x|j)$ will become so large such that the discrimination between the different motion classes will be impossible.

5 Classification Experiments

We now benefit from the fact that it is no longer necessary to track the outline of the person in order to classify the type of motion. This allows us to test the newly developed method on examples which are considerably more complex than these used in section 3. The first data set contains sequences of four different people running, skipping and walking. Due to the low contrast between foreground and background it would be difficult to track the outline of the people using an edge-based contour tracker. The second set of test sequences consists of a set of aerobics exercises similar to those used previously. But here we include the star jump as third motion class.

All sequences of people walking were recorded at the same time and place. Images of the people involved in the experiment are shown in figure 9. The group of people contains three males and one female all between 20 and 30 years old. In all the data set includes 77 sequences whose lengths vary between 2 sec. and 5 sec.. A simple blob tracker based on motion detectors is applied to locate the foreground window in every frame.

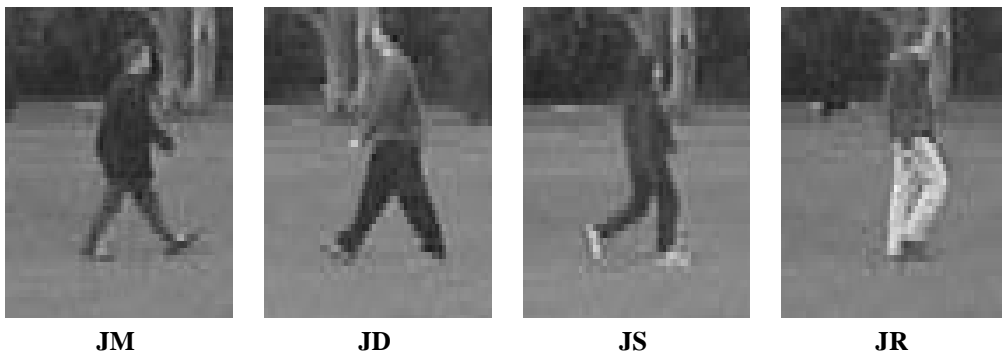


Figure 9: **People involved in the experiment.** Shown are the four people from whom sequences of walking, skipping, and running were recorded. The group contains three males and one female person. It can be seen that the four people are of different height.

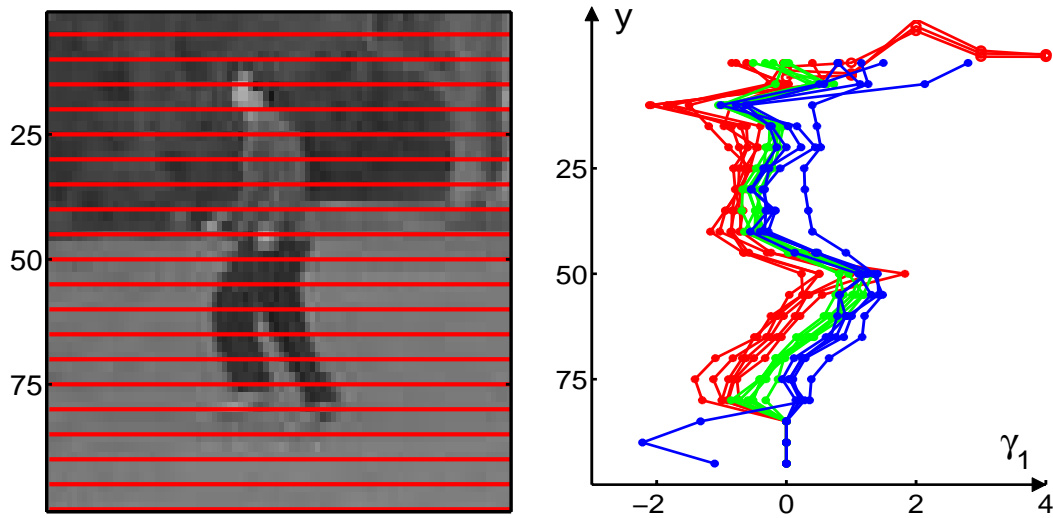


Figure 10: **Skew vectors $\gamma_1(I)$ for different walking styles.** The skew vectors $\gamma_1(I)$ (4) are computed for a set of 20 epipolar slices. The exact position of the slices are indicated as red lines in the image on the left. The skew vectors, $\gamma_1(I)$, for each of the sequences are presented in the graph on the right. The skew vectors are colour coded. The colour red corresponds to running, green to skipping, and blue to walking. It can be observed that the skew factors of top and bottom slices are implausible. This is a consequence of the naive threshold set in (see equation (3)). The amount of variation in the upper half of the body can be explained by the body motion with respect to the centroid of the body. As a result of computing the skew factors of the different epipolar slices the different walking styles are ordered according to their average speed. In an experiment a linear discrimination analysis of the skew vectors $\gamma_1(I)$ for a set of four different people performing running, skipping and walking was used to demonstrate that the three different motion classes can be separated.

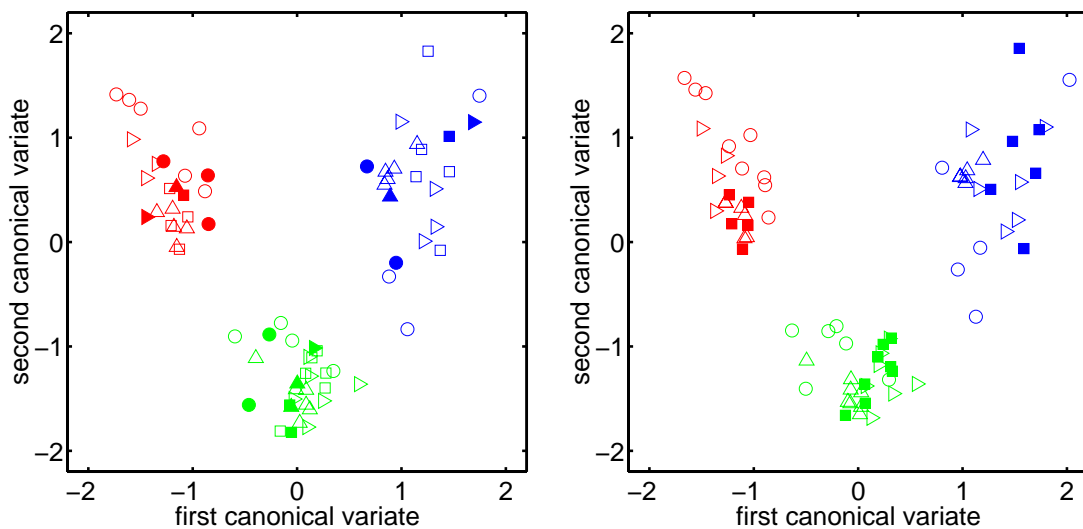


Figure 11: **Classification of walking styles.** The set of sequences containing running, skipping and walking of four different people was divided into a training and a test set. A linear fisher discriminant was then computed for the training data. The test data was mapped onto the first and second canonical variates. Every person in the training set is represented by one of the following symbols (\triangle - JM, \square - JS, \triangleright - JR, \circ - JD). The red colour indicates running, green skipping and blue walking. The samples belonging to the training set are shown as outlines whereas the samples of the test set are shown as filled in symbols. Two trials were conducted. The graph on the left shows the results obtained from the first trial. Here the training set contained samples of each person and all three motion types. In a second trial the training set only contained the data of three people performing running, skipping and walking. Although no normalisation for the height of the person is made the methods allows to reliably detect the type of walking style.

In order to visualise some of the results the resulting skew vectors of one person, $\gamma_1(I)$, are presented in figure 10. A total number of 20 epipolar slices are analysed for each of the sequences. As a result of the naive threshold C (3) the skew factors for the top and bottom slices are implausible. The number of locations (x, t) where the modulus of the gradient $|\nabla I(x, t)|$ exceeds the threshold C is small hence the estimate of the distribution of $|\alpha|$ is noisy and unreliable. For that reason the top and bottom slices are discarded. It should be noted that the vectors of skew factors shown in figure 10 give rise to a very natural interpretation of the characteristics of the three different motion types. This method orders these different walking styles in some kind of continuum where the two extremes are defined by running and walking. Here we of course make use of the fact that the walking motions are periodic. In the next section we will demonstrate that the same method can be used to recognise non-repetitive motions when very short time intervals are used.

The question of how well the skew vectors discriminate the different motions was tested in a second experiment. The total number of sequences was divided into a training set and a test set. The first and second canonical variates, also known as the first two Fisher linear discriminants, [21] were computed from the training data. One problem is, of course, that the number of data points compared to the dimensionality of the feature space is small. In order to avoid over-fitting, the training set is enlarged by a number of random samples. The samples for each class are drawn from a multivariate Gaussian distribution whose center and diagonal covariance matrix are estimated from the training data. The test data was subsequently mapped onto these canonical variates. In a first trial the training and the test set contained samples of all four people and all three motions. In a second trial the training set contained samples of all three motions but only of three people. Both results are shown in figure 11. Without applying any normalisation with respect to the height of the person the motion classes are well separated and allow a reliable classification of the class of walking style. The second trial was repeated such that every person was eliminated from the training set. In all experiments we are able to separate the classes of motion.

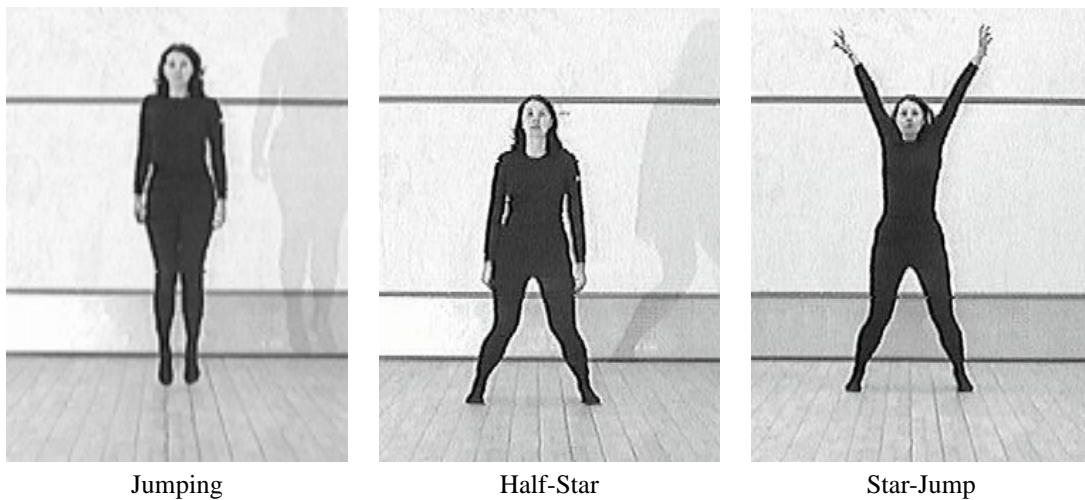


Figure 12: **Aerobics motions.** The figure shows single frames taken from the test sequences containing aerobics exercises. The sequences containing jumping and the half star motion were already used for the automatic segmentation experiments in section 3. The full Star-Jump or jumping jack was not used because the contour tracker failed to track the arm motion correctly. Because this analysis does not require any contour tracking this problem is now eliminated.

The second set of test sequences contains three different gymnastic exercises: Jump, Half-Star jump and Star-Jump (or jumping jack) (see figure 12). These sequences do not require any preprocessing since the exercise is performed on the spot.

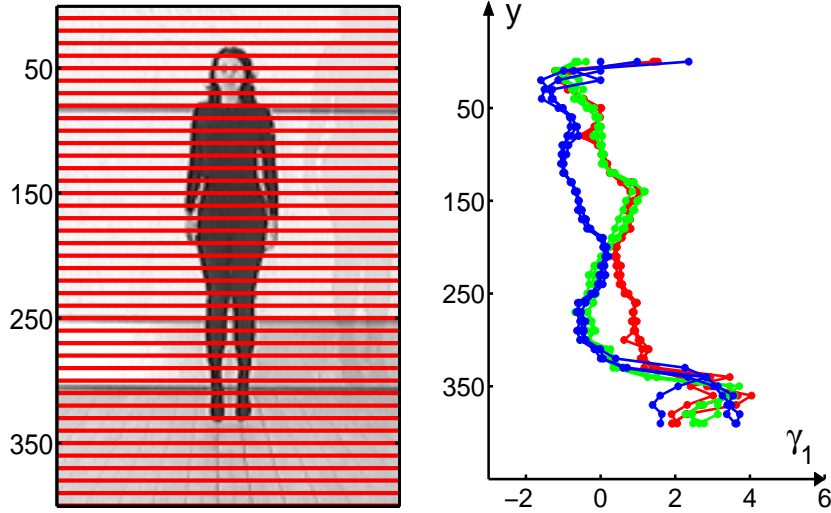


Figure 13: **Skew vectors $\gamma_1(I)$ for the aerobic motions.** In these experiments the skew factors $\gamma_1(I)$ (see equation (4)) are computed for a set of 40 epipolar slices. The exact position of the slices are again indicated as red lines in the image on the left. The skew vectors $\gamma_1(I)$ for each of the test sequences are presented in the graph in the right. The skew vectors are colour coded. The colour red corresponds to Jumping, green to Half-Star jump and blue to the full Star-Jump. It should be noted that although this method is more suitable for analysing leg motion (see text) the presence of arm motion is detected correctly. And moreover the fact that the skew factors are a relatively smooth function with respect to y indicate that the estimates are not particularly noisy.

As it was mentioned in the introduction to this section it is now no longer necessary to track the contour of the person. The skew factors $\gamma_1(I)$ can be estimated directly from the image sequence. It needs to be noted that the method of analysing epipolar slices is particularly well suited to analysing the leg motion. This is because of the length of the legs and their relative angle with the epipolar slice. Arm motion does not give rise to a continuous motion pattern which can be observed in the epipolar slices. But the resulting skew vectors, shown in figure 13 clearly reflect the presence of arm motion. Similar to the skew factors for the walking sequences a very natural interpretation of the data is obtained. Since the leg motion of the Half-Star and the Star-Jump is identical, the vector of skew factors are similar for the slices corresponding to the lower half on the body. A similar effect can be observed for the lack of arm motion in the other two motions in the upper body. Needless to say a linear projection of the skew factors onto the first and second canonical variates shows that the data is well separated. Hence this is a suitable feature set for classification. Figures 13 and 10 both indicate that the skew factors $\gamma_1(I, y)$ are continuous with respect to y . This results, of course, from the fact that humans move smoothly. This fact will be used later, when we estimate the skew vectors $\gamma_1(I)$ for small samples sizes.

6 Automatic Segmentations

Many applications require the instantaneous classification of the type of motion. The skew vectors $\gamma_1(I)$ are computed for a set of consecutive frames or in other words, of a spatio-temporal cube of a certain temporal length T . Apart of the noise of the measurement model itself, the distribution of the skew vectors $\gamma_1(I)$ for each class of motion will now depend on the length of the temporal window. This is due to the phase dependency of the motion itself. In a first experiment we attempt to classify

each temporal window using a very basic model for the class densities. Like in the computation of the linear discriminant analysis in the previous section, the distributions of the skew vectors $\gamma_1(I)$ for each motion class are modelled as multivariate Gaussians with mean μ and diagonal covariance Σ . The Mahalanobis distance between the skew vectors $\gamma_1(I)$, computed for each interval of length T , and the learnt centers μ_i , is used to classify the observed motion.

We believe, that it is important to test the use of the skew vectors $\gamma_1(I)$ using a very basic classification technique. Unlike Hidden Markov Models, such an approach discards, of course, the temporal information about the mean duration of each motion. But parameters for models with hidden parameters are usually learnt using some form of expectation maximisation learning rule. These algorithms do not guarantee to find a local minimum and depend heavily on an initialisation step. If the results obtained by a basic classifier are promising it will be possible to refine them using a more elaborate model.

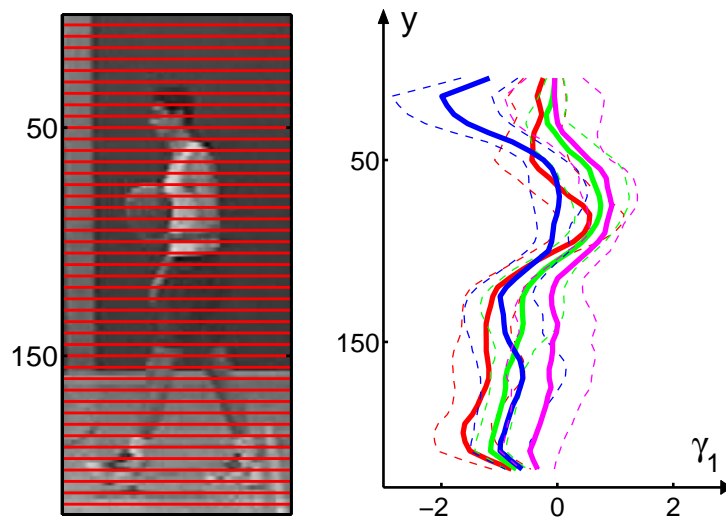


Figure 14: **Skew vectors $\gamma_1(I)$ for basketball player.** The skew vectors $\gamma_1(I)$ (4) are computed for a set of 44 epipolar slices. The feature vectors are convolved with a smoothing filter to reduce the measurement noise (see text). As before the left image shows the location of the epipolar slices. The distributions of the skew vectors of each motion class is modelled by a multivariate Gaussian with a diagonal covariance matrix. The graph on the right shows the mean and standard deviation of each motion class. Four motion classes are used in this experiment: running (in red), walking (in green), turning (in magenta), and throwing (in blue). It can be easily observed that the throw is the only motion that involves a considerable amount of arm motion. The turning motion involves very little arm and leg motion. The motion classes running and walking, however, show some overlap.

7 Segmentation Results

Two sets of image sequences were analysed. The first set contains three sequences of a mixture of aerobics exercises shown in figure 12. The temporal length of the spatio-temporal cube T is chosen to be 25 fields or half a second. Like in the experiment shown in figure 13, a number of 40 epipolar slices are used to compute $\gamma_1(I)$. The distributions of the skew vector for each class Jumping, Half-Star, and Star-Jump are learnt from a separate set of training sequences each of which only contains one type of motion. In total, 114 half-second intervals were analysed. Out of these 8 were classified wrongly, which corresponds to an error rate of 7%.

These classification errors are due to measurement noise. In order to reduce this measurement noise we convolve each feature vector with a smoothing filter. This is justified by the following observations. Firstly, we expect the true skew factors $\gamma_1(I, y)$, as mentioned before, to be continuous with respect to y . Secondly, the noise of the skew estimate on two adjacent slices $\gamma_1(I, y_i)$ and $\gamma_1(I, y_j)$ can be assumed to be independent. The reason being that the support of the filters, $\delta_t\varphi$ and $\delta_x\varphi$, needed to compute α (see equation 3), on the two adjacent slices are disjoint. As a result the number of misclassifications is reduced to 3 out of 114 samples. This corresponds to an error rate of 2.6 %.

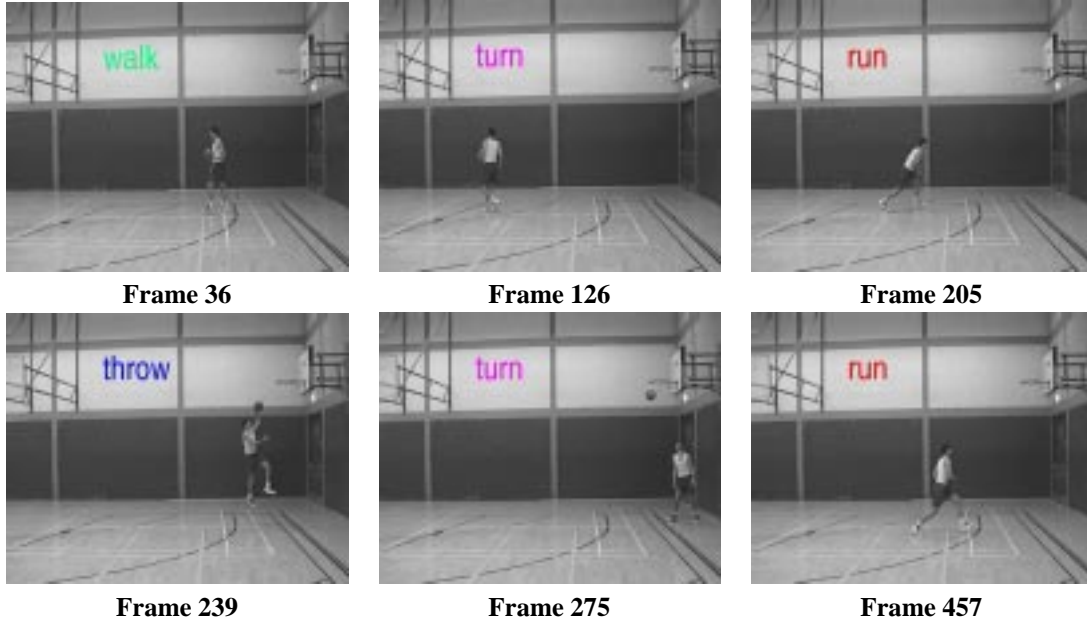


Figure 15: **Automatically annotated basketball sequence.** Shown are single frames out of the automatically annotated sequence. The person runs towards the basket dribbling the ball. The ball is then thrown into the basket. After the ball is caught the person walks back. All of these stages are identified correctly. Only the decision between running and walking is sometimes ambiguous.

Finally we analyse sequences of a person playing basketball. Two sequences, each of which is 50 seconds length, are used. Like in the previous example the length of the spatio-temporal cube T is set to be 25 fields of half a second and the skew vectors $\gamma_1(I)$ were again convolved with a smoothing kernel. The foreground region was tracked using a simple blob tracker. The training data is obtained by labelling each half second interval in one of the sequences as belonging to one of the following motion classes: walking, running, turning, and throwing. As before, each of these motion classes is modelled by a multivariate Gaussian with diagonal covariance matrix. The resulting means for each of the motion classes are shown in figure 14. As it can be seen in the figure the distributions of walking and running have some overlap. As a result walking is occasionally misclassified as running. In total, 10 out of 99 samples are misclassified which corresponds to a misclassification rate of 10 %. All remaining classes of motion are identified correctly. Single frames of the sequence are shown in figure 15. Taking the difficulty of the sequence into account this is a very promising result. Potentially this error rate results from the fact that during training some intervals which contain a transition from standing still to running were labelled as running. Assuming that the camera is stationary this error rate can easily be reduced by taking prior information, such as the output from the blob tracker, into account.

8 Conclusion

We have demonstrated that a particle filter with mixed states can be used for classifying motions online. In the approach presented by [5] the discrete states are used up for modelling atomic motions. In that context a bi-level recursive algorithm [20] could be tried for automatic segmentation. But these algorithms are notoriously computationally expensive. Here we show that a single autoregressive model is a serious candidate as a model of atomic motions. This leaves the discrete state free for classification. A Markov chain is used to model the state transitions and to apply long-term continuity constraints. It was demonstrated that the Markov chain can be applied more effectively by using partial importance sampling. Partial importance sampling does indeed improve the asymptotical efficiency, but it does not seem to be capable of reducing the asymptotic error rate.

In order to address the problem of tracker initialisation and foreground modelling we present a novel feature extraction method for the classification of human motion. The method produces feature vectors which can easily be interpreted by inspection. The strength of this feature extraction method is demonstrated for both the classification of sequences containing only one type of motion, as well as, the automatic segmentation of sequences containing mixed motions. The experiments on classifying different walking styles show, that this method is able to generalise from the training data, i.e. it is possible to classify the motion of a person who was not included in the training set. Although we have successfully eliminated problems related to the contour tracking approach there are still a number of open problems. One important question which needs to be addressed is how this method can be used to obtain a view-point independent classification. Yacoob and Black [29] addressed this problem by defining a similarity measure which is invariant under a certain transformation set. Also the effect of background motion need to be investigated in more detail.

References

- [1] A. Baumberg and D. Hogg. Learning flexible models from image sequences. In *Proc. 3rd European Conf. Computer Vision, Stockholm, Sweden*, pages 299–308. Springer-Verlag, 1994.
- [2] M. J. Black and A. D. Jepson. A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. In *Proc. 5th European Conf. Computer Vision, Freiburg, Germany*, pages 909–924, 1998.
- [3] A. Blake and M. Isard. *Active Contours*. Springer, 1998.
- [4] R. C. Bolles and H. H. Baker. Epipolar-plane image analysis: A technique for analyzing sequences. In *Workshop on Computer Vision, Representation and Control, Shanty Creek, MI*, pages 168–178, October 1985.
- [5] C. Bregler. Learning and recognizing human dynamics in video sequences. In *Proc 11th IEEE Computer Vision and Pattern Recognition, San Jaun, PR*, pages 568–574, 1997.
- [6] O. Chomat, J. Martin, and J.L. Crowley. A probabilistic sensor for the perception and recognition of activities. In *Proc. 6th European Conf. Computer Vision, Dublin, Ireland*, pages 487–503, 2000.
- [7] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *IEEE Computer Vision and Pattern Recognition, Hilton Head, SC*, pages 126–133, 2000.

- [8] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEEE Proc.*, 140(2):107–113, 1993.
- [9] J.M. Hammersley and D.C. Handscomb. *Monte Carlo methods*. Methuen, 1964.
- [10] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. 4th European Conf. Computer Vision, Cambridge, UK*, volume 1, pages 343–356, 1996.
- [11] M. Isard and A. Blake. Condensation – conditional den-sity propagation for visual tracking. *Int. J. Computer Vision*, 28(1):5–28, 1998.
- [12] M. Isard and A. Blake. A mixed-state condensation tracker with automatic model switching. In *Proc. 6th Int. Conf. on Computer Vision, Bombay, India*, pages 107–112, 1998.
- [13] M.A. Isard and A. Blake. ICondensation: Unifying low-level and high-level tracking in a stochastic framework. In *Proc. 5th European Conf. Computer Vision, Freiburg, Germany*, volume 1, pages 893–908, 1998.
- [14] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 1991.
- [15] G. Kitagawa. Monte Carlo filter und smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5:1–25, 1996.
- [16] H. Lütkepohl. *Introduction to Multiple Time Series Analysis (second edition)*. Springer, 1993.
- [17] S. A. Niyogi and E. H. Adelson. Analyzing and recognizing walking figures in xyt. In *Proc 9th IEEE Computer Vision and Pattern Recognition, Seattle, WA*, pages 469–474, 1994.
- [18] S. A. Niyogi and E. H. Adelson. Analyzing gait with spatiotemporal surfaces. In *Workshop on Non-Ridgit Motion and Articulated Objects, Austin, Texas*, 1994.
- [19] B. North, A. Blake, M. Isard, and J. Rittscher. Learning and classification of complex dynamics. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(9):1016–1034, 2000.
- [20] L. Rabiner and J. Bing-Hwang. *Fundamentals of speech recognition*. Prentice-Hall, 1993.
- [21] B. R. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.
- [22] H. Scheffé. *The Analysis of Variance*. J. Wiley & Sons, Inc, 1959.
- [23] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2D image motion. In *Proc. 6th European Conf. Computer Vision, Dublin, Ireland*, volume 2, pages 702–718, 2000.
- [24] T. Starner and A. Pentland. Visual recognition of American Sign Language using Hidden Markov Models. In *Proc 1st Int. Conf. on Automatic Face and Gesture Recognition, Killington, VT*, 1995.
- [25] J. Sullivan, A. Blake, M. Isard, and J. MacCormick. Object localisation by Baysian correlation. In *Proc. 7th Int. Conf. on Computer Vision, Corfu, Greece*, volume 2, pages 1068–1075, 1999.

- [26] J. Sullivan, A. Blake, and J. Rittscher. Statistical foreground modelling for object localisation. In *Proc. 6th European Conf. Computer Vision, Dublin, Ireland*, volume 2, pages 307–323, 2000.
- [27] A. D. Wilson and A. F. Bobick. Nonlinear parametric Hidden Markov Models. Technical report, MIT Media Lab Perceptual Computing Group Technical, Report 424, Massachusetts Institute of Technology, 1997.
- [28] A. D. Wilson and A. F. Bobick. Recognition and interpretation of parametric gesture. In *Proc 11th IEEE Computer Vision and Pattern Recognition, San Juan, PR*, pages 329–336, 1997.
- [29] Y. Yacoob and M. J. Black. Parameterized modelling and recognition of activities. In *Proc. 6th Int. Conf. on Computer Vision, Bombay, India*, 1998.