Int J Comput Vis (2009) 83: 121–134
DOI 10.1007/s11263-008-0158-0

Learning Generative Models for Multi-Activity Body Pose Estimation

Tobias Jaeggli · Esther Koller-Meier · Luc Van Gool

Received: 30 January 2008 / Accepted: 11 July 2008 / Published online: 31 July 2008

Abstract We present a method to simultaneously estimate 3D body pose and action categories from monocular video sequences. Our approach learns a generative model of the relationship of body pose and image appearance using a sparse kernel regressor. Body poses are modelled on a low-dimensional manifold obtained by Locally Linear Embedding dimensionality reduction. In addition, we learn a prior model of likely body poses and a dynamical model in this pose manifold. Sparse kernel regressors capture the nonlinearities of this mapping efficiently. Within a Recursive Bayesian Sampling framework, the potentially multimodal posterior probability distributions can then be inferred. An activity-switching mechanism based on learned transfer functions allows for inference of the performed activity class, along with the estimation of body pose and 2D image location of the subject. Using a rough foreground segmentation, we compare Binary PCA and distance transforms to encode the appearance. As a postprocessing step, the globally optimal trajectory through the entire sequence is estimated, yielding a single pose estimate per frame that is consistent throughout the sequence. We evaluate the algorithm on challenging sequences with subjects that are alternating between running and walking movements. Our experiments show how the dynamical model helps to track through poorly segmented low-resolution image sequences where tracking otherwise fails, while at the same time reliably classifying the activity type.

Keywords Monocular pose estimation · Machine learning · Dimensionality reduction · Activity recognition · Human locomotion

T. Jaeggli (✉) · E. Koller-Meier · L. Van Gool
ETH Zurich, Zurich, Switzerland
e-mail: jaeggli@vision.ee.ethz.ch

E. Koller-Meier
e-mail: ebmeier@vision.ee.ethz.ch

L. Van Gool
e-mail: vangool@vision.ee.ethz.ch

L. Van Gool
KU Leuven, Leuven, Belgium

1 Introduction

Monocular body pose estimation is difficult, because a certain input image can often be interpreted in different ways. Image features computed from the silhouette of the tracked figure hold rich information about the body pose, but silhouettes are inherently ambiguous, *e.g.* due to the Necker reversal. Through the use of prior models this problem can be alleviated to a certain degree, but in many cases the interpretation is ambiguous and multi-valued throughout the sequence.

Several approaches have been proposed to tackle this problem, they can be divided into *discriminative* and *generative* methods. Discriminative approaches directly infer body poses given an appearance descriptor, whereas generative approaches provide a mechanism to predict the appearance features given a pose hypothesis, which is then used in a generative inference framework such as particle filtering or numerical optimisation.

Recently, statistical methods have been introduced that can learn the relationship of pose and appearance from a training data set. They often follow a discriminative approach and have to deal explicitly with the nonfunctional nature of the multi-valued mapping from appearance to pose (Rosales and Sclaroff 2001; Thayananthan et al. 2006;

Sminchisescu et al. 2005; Agarwal and Triggs 2005; Jaeggli et al. 2006). Generative approaches on the other hand typically use hand crafted geometric body models to predict image appearances (*e.g.* Sidenbladh et al. 2000, see Forsyth et al. 2006; Moeslund et al. 2006 for an overview).

We propose to combine the generative methodology with a learning based statistical approach. The mapping from pose to appearance is single-valued and can thus be seen as a nonlinear regression problem. We approximate the mapping with a Relevance Vector Machine (RVM) kernel regressor (Tipping 2000) that is efficient due to its sparsity. See Fig. 1 for an illustration of the one-to-many regression problem. Although single-valued, the appearance prediction will be subject to uncertainty, because other factors than just the body configuration (pose) may affect appearance (clothing, physical constitution, lighting conditions *etc.*). This is taken into account by learning the prediction variances of the mapping.

The observations are available in the form of roughly segmented monocular image sequences that are obtained by a pre-processing step such as motion segmentation, background subtraction or other. A main focus of the proposed approach lies on the ambiguities and uncertainties that are inherent in body tracking from such input. Recursive Bayesian Sampling (Isard and Blake 1998a; Doucet et al. 2000a) offers a framework for dealing with non-Gaussian and multimodal body pose posteriors and allows us to integrate the nonlinear learned dynamical model. However, sampling-based algorithms are generally not applicable for inference in high-dimensional state spaces like the space of body poses. We therefore use Locally Linear Embedding (LLE, Roweis and Saul 2000) to find a low-dimensional embedding of our 60-dimensional pose parametrisation. With 4 LLE dimensions, the considered motions can be captured reasonably well.

The tasks of body pose estimation and activity recognition are strongly related. While a sequence of inferred body poses might be used for activity recognition, a known activity class can also help the pose estimation, *e.g.* by selecting an appropriate context specific prior. The proposed method estimates 3D body pose and action categories simultaneously. We learn strong dimensionality-reduced models of feasible body poses that belong to a certain activity or motion pattern, as well as the temporal evolution of the body poses over time. Furthermore, the transition functions between different activities are learned from training data too. In this article we investigate typical human motion patterns such as walking and running. Rather than learning a unified representation that contains both walking and running motions, we learn separate activity specific models that allow us to explicitly recognise the performed activity along with the pose estimation, using a switching mechanism of the inference algorithm.

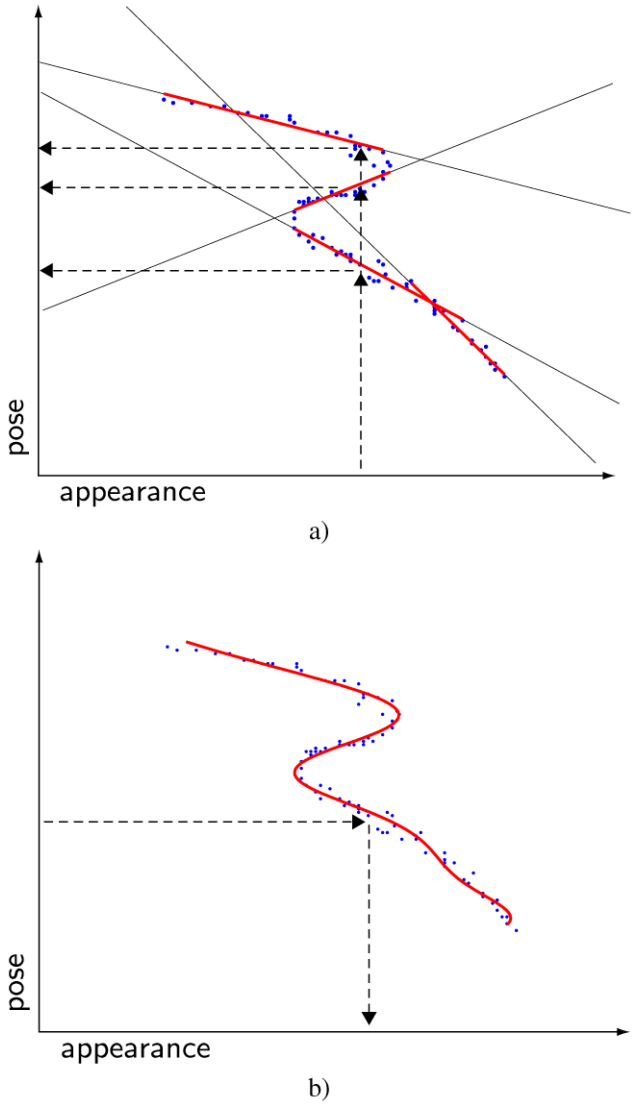

Fig. 1 Illustration of a discriminative one-to-many mapping with a mixture of linear regressors (**a**), and of a generative mapping from pose space to appearance space with a single nonlinear regressor (**b**)

The main contributions of this article are the generative appearance modelling, the tracking in a LLE-reduced pose representation with a nonlinear dynamical model, simultaneous recognition of multiple action categories, and the extraction of a globally optimal trajectory through the entire sequence.

2 Related Work

There is a wide variety of literature about body pose estimation and tracking (see Forsyth et al. 2006 for an overview). Here we will have a look at the application of statistical methods that infer poses from one or multiple camera streams. Many authors adopt a discriminative strategy to

infer poses directly from image descriptors (Rosales and Sclaroff 2001; Thayananthan et al. 2006; Sminchisescu et al. 2005; Agarwal and Triggs 2004a, 2005; Grauman et al. 2003; Sun et al. 2006).

Synchronous image sequences from multiple cameras typically provide enough information to resolve ambiguities. The discriminative mapping from descriptors to body poses can thus be modelled using a *single* regressor. In Sun et al. (2006), a new image descriptor is introduced based on a voxel representation that is derived from segmented images of multiple cameras. This descriptor can then be directly mapped into pose space. In Grauman et al. (2003) multiple silhouette image descriptors and corresponding pose descriptors are concatenated and modelled with a mixture of Probabilistic PCA; poses can then be inferred given multiple views of the subject.

Monocular approaches have to deal with the one-to-many discriminative mapping from appearance to pose. This issue is explicitly addressed in Rosales and Sclaroff (2001), Thayananthan et al. (2006), Sminchisescu et al. (2005), Agarwal and Triggs (2005), Jaeggli et al. (2006) by learning *multiple* mappings in parallel as a mixture of regressors. In order to choose between the different hypotheses that the different regressors deliver, Rosales and Sclaroff (2001), Thayananthan et al. (2006) use a geometric model that is projected into the image to verify the hypotheses. Inference is performed for each frame independently in Rosales and Sclaroff (2001). In Thayananthan et al. (2006) a temporal model is included using a bank of Kalman filters, and a Viterbi algorithm finds a path through the peaks of the posterior distribution. In Sminchisescu et al. (2005), Agarwal and Triggs (2005), Jaeggli et al. (2006) gating functions are learned along with the regressors in order to pick the right regressor(s) for a given appearance descriptor. The distribution is propagated analytically in Sminchisescu et al. (2005), and temporal aspects are included in the learned discriminative mapping, whereas Agarwal and Triggs (2005) adopts a generative sampling-based tracking algorithm with a first-order autoregressive dynamic model. In Jaeggli et al. (2006) discriminative analytical inference and generative sample-based inference are combined in a Rao-Black wellised particle filter. This allows for efficient inference in the high dimensional pose space in combination with the non-parametric posterior distributions that occur when the 2D image location has to be inferred by the algorithm as well.

The mentioned purely discriminative approaches work in a bottom-up fashion, starting with the computation of the image descriptor, which requires the location of the figure in the images to be known beforehand. When including 2D bounding box estimation in the tracking problem, a learned dynamical model of the appearance might help the bounding box tracking, and avoid losing the subject when it is temporarily occluded. To this end, Lim et al. (2006) learns a subject-specific dynamic appearance model from a small set of initial frames, consisting of a low-dimensional embedding of the appearances and a motion model. This model is used to predict the location and appearance of the figure in future frames, within a *CONDENSATION* tracking framework. Similarly, low-dimensional embeddings of appearance (silhouette) manifolds are found using LLE in Elgammal and Lee (2004), where additionally the mapping from the appearance manifold to 3D pose in body joint space is learned using radial basis function (RBF) interpolants, allowing for pose inference from sequences of silhouettes.

Instead of modelling manifolds in *appearance* space, Wang et al. (2006), Sminchisescu and Jepson (2004), Li et al. (2006) work with low dimensional embeddings of body *poses*. In Wang et al. (2006), Urtasun et al. (2006), the low-dimensional pose representation, its dynamics, and the mapping back to the original pose space are learned in a unified framework. This approach does not include a learned statistical model of image appearance. Our method also models *pose* manifolds rather than *appearance* manifolds, because the pose manifold has fewer self-in tersections than the appearance manifold, making the dynamics and tracking less ambiguous.

To model the nonlinear dynamics of human motion, different approaches have been proposed. In Pavlovic et al. (2001), Agarwal and Triggs (2004b), Li et al. (2007) mixtures of linear autoregressive motion models (respectively piecewise linear models) were used, where the inference algorithm switches between a number of discrete states corresponding to the different linear models. These models are learned using EM-like optimisation. Our method is in line with Wang et al. (2006), Urtasun et al. (2006), Lee and Elgammal (2007), where temporal predictions are obtained using nonlinear regression. In such a way, smooth motion flow fields over the entire pose space can be learned directly. In contrast to a piecewise linear model, there is thus no need to learn a finite number of states that have no semantic meaning that is of interest for our task, and in particular we don't have to select the optimal number of mixture components.

Most related to our approach are two recent publications that also model the appearance of moving persons in a generative way. In Lee and Elgammal (2007) a low dimensional embedding of the kinematics is obtained using joint angles data. View-based nonlinear mapping functions from the kinematic manifolds to multi-view appearance are used to infer poses in a Bayesian tracking framework. They model the view dependencies of the appearance on a low-dimensional posture-independent view manifold. In Navaratnam et al. (2007) a shared low-di mensional latent space with nonlinear mappings into the pose and appearance spaces is learned using an extension of the GP-LVM (Lawrence 2005)

framework. The graphical structure of the learned model is thus similar to the method proposed in this article, with the exception of the dynamics, which are not explicitly modelled in Navaratnam et al. (2007). Furthermore the authors propose the use of unlabelled data to improve the learned model.

Regarding activity switching, Isard and Blake (1998b) have proposed a state switching mechanism, where different dynamical models are chosen, depending on a discrete state variable. In our approach, the different states (activities) involve separate models for pose, dynamics and appearance.

Our approach differs from the above-mentioned papers in that it simultaneously tracks in a state space that includes body pose, 2D bounding box location and a discrete activity label. Furthermore, we present a full-fledged pipeline with *generative* rather than *discriminative* modelling of the appearance, entirely based on learned models. The framework is built-up in a module based manner. Some choices of precise statistical methods that are applied for the individual modules are mainly based on practical considerations (*e.g.* efficiency, sparsity). They could be substituted by equivalent methods, like *e.g.* Isomap (Tenenbaum et al. 2000) instead of LLE, and regularised kernel regressors or Gaussian processes instead of RVMs.

The remainder of this paper is organised as follows. Section 3 introduces our learned models. In Sect. 4 the sample-based inference algorithm is presented and Sect. 5 shows experimental results on different video sequences.

3 Statistical Modelling

Figure 2a shows an overview of the tracking framework, reduced to a single activity category for clarity. Central element is the low-dimensional body pose parametrisation, with learned mappings back to the original pose space and into the appearance space. In this section all elements of the framework will be described in detail.

Our models were trained on motion capture data sets of different subjects, running and walking at different speeds. Walking and running training examples were separately processed to train activity specific models.

3.1 Pose and Motion Model

Representations for the full body pose configuration are high dimensional by nature; our current representation is based on 3D joint locations of 20 body locations such as hips, knees and ankles (see Fig. 2b, but any other representation (*e.g.* based on relative orientations between neighbouring limbs) can easily be plugged into the framework. To alleviate the difficulties of high dimensionality in both the learning and inference stages, a dimensionality reduction step identifies a low dimensional embedding of the body pose representations. We use Locally Linear Embedding (LLE) (Roweis and Saul 2000), which approximately maintains the local neighbourhood relationships of each data point and allows for global deformations (*e.g.* unrolling) of the dataset/manifold.

LLE dimensionality reduction is performed on all poses in the data set that belong to a certain activity, and expresses

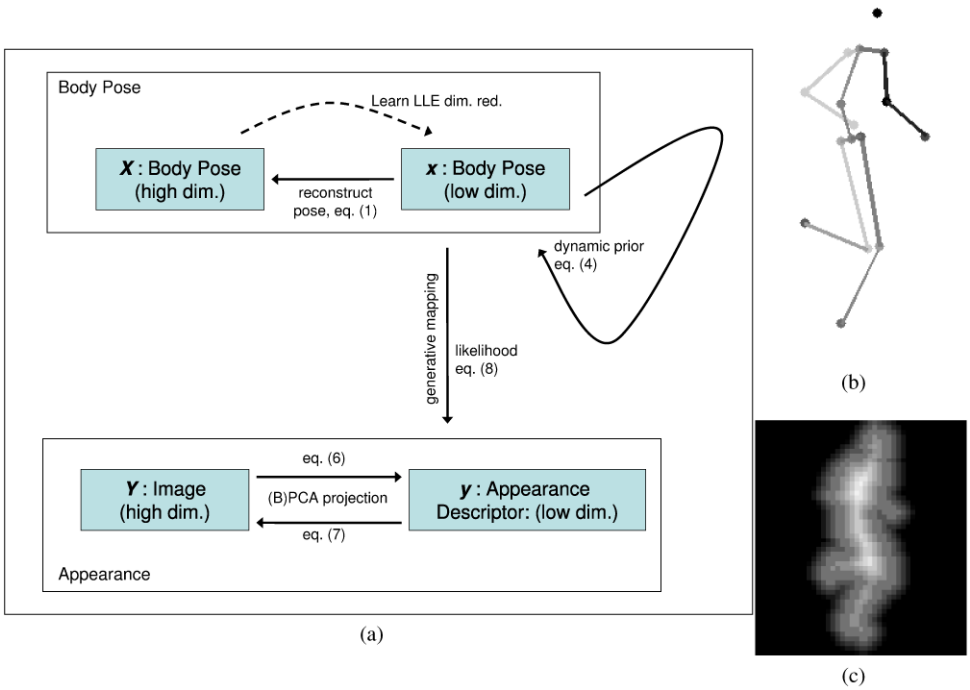

Fig. 2 (**a**) An overview of the tracking framework. *Solid arrows* represent signal flow during inference, the *dashed arrow* stands for the nonlinear dimensionality reduction during training. The figure refers to equations in Sect. 3. (**b**) Body pose representation as a number of 3D joint locations. (**c**) Distance transformed image descriptor $dt(Y)$. Each pixel value is proportional to the distance to the silhouette, and its sign indicates whether the pixel lies inside the silhouette

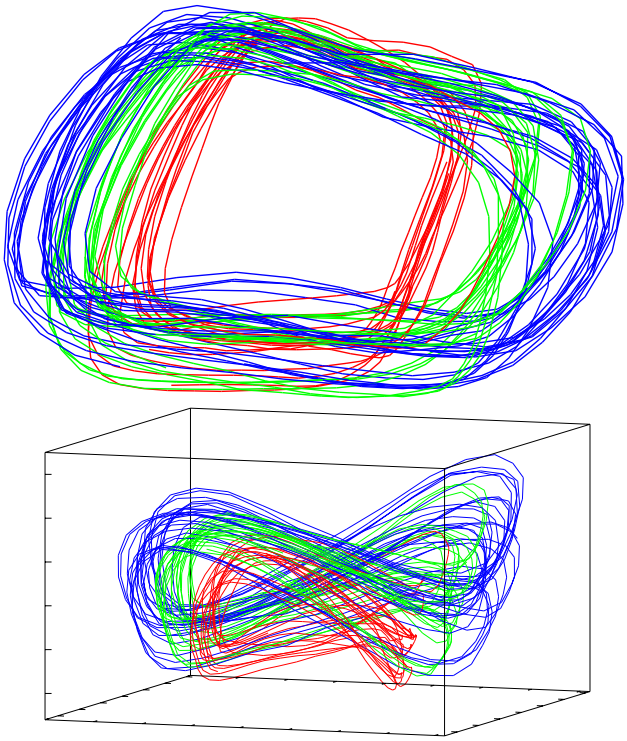

Fig. 3 (Color online) Low-dimensional manifold of walking data obtained by Locally Linear Embedding. Three dimensions of the four-dimensional representation are visualised here, from two different views. The different colours indicate different walking speeds (*red*: 2.5 km/h, *green*: 4.2 km/h, *blue*: 6 km/h)

each data point in a space of desired low dimensionality (see also Fig. 3). However, LLE does not provide explicit mappings between the two spaces, that would allow to project new data points (that were not contained in the original data set) between them. Therefore, we model the reconstruction projection from the low-dimensional LLE space to the original pose space with a kernel regressor.

$$X = f_p(x) = W_p \Phi_p(x) \tag{1}$$

Here, X and x are the body pose representations in original resp. LLE-reduced spaces, Φ_p is a vector of kernel functions, and W_p is a sparse matrix of weights, which are learned with a Relevance Vector Machine. We use Gaussian kernel functions, computed at the training data locations. Separate models are learned for the two distinct activities, $f_p^w(x^w)$ and $f_p^r(x^r)$. In the following we will use superscripts (*e.g.* w for $walk$ and r for run) to indicate activity categories in the notation if necessary and omit them if the same formulation holds for all actions.

The training examples form a periodic twisted 'ring' in LLE space, with a curvature that varies with the phase within the periodic movement. A linear dynamical model, as often used in tracking applications, is not suitable to predict future poses on this curved manifold. We view the nonlinear dynamics as a regression problem, and model it using another RVM regressor, yielding the following *dynamic* prior,

$$p_d(x_t|x_{t-1}) = \mathcal{N}(x_t; x_{t-1} + f_d(x_{t-1})\Delta_T, \Sigma_d), \tag{2}$$

where $f_d(x_{t-1}) = W_d \Phi_d(x_{t-1})$ is the nonlinear mapping from poses to local velocities in LLE pose space, Δ_T is the time interval between the subsequent discrete timesteps $t-1$ and t, and Σ_d is the variance of the prediction errors of the mapping, computed on a hold-out data set that was not used for the estimation of the mapping itself. Again, the dynamics are learned separately for the different action categories.

Not all body poses that can be expressed using the LLE pose parametrisation do correspond to valid body configurations that can be reached with a human body. The motion model described so far does only include information about the temporal evolution of the pose, but no information about how likely a certain body pose is to occur in general. In other words, it does not yet provide any means to restrict our tracking to feasible body poses. The additional prior knowledge about feasible body poses, or likely poses for a given activity, is introduced as a *static* prior that is modelled with a Gaussian Mixture Model (GMM),

$$p_s(x) = \sum_c^C p_c \mathcal{N}(x; \mu_c, \Sigma_c), \tag{3}$$

with C the number of mixture components and p_c, μ_c and Σ_c the mixture proportions and parameters of the Gaussian components. The influence of this pose prior can be kept low, avoiding a distortion of the tracking results towards typical average motion. We introduce a weighting factor $\lambda > 1$ and obtain the following formulation for the temporal prior by combination with the *dynamic* prior $p_d(x_t|x_{t-1})$.

$$p(x_t|x_{t-1}) \propto p_d(x_t|x_{t-1})\ p_s(x_t)^{\frac{1}{\lambda}} \tag{4}$$

We also want to model the transition between the considered action categories, that each have their own low dimensional pose parametrisation expressed in distinct LLE spaces. Informally, we want to find walking poses that are very similar to a given running pose and vice versa, since we know that the transition is performed smoothly, without any sudden or jerky 'jump' of the body configuration.

Given our distinct training sets of walking and running poses, two sets of training pairs are generated by looking for the most similar running pose for every walking pose and vice versa, and the nonlinear mapping between these pairs is modelled using two sparse kernel regressors $f_{switch}^{r\to w}(x^r)$ and $f_{switch}^{w\to r}(x^w)$. This can be generalised to more action categories[1] and leads to the following motion model, where the

[1]The number of transitions grows quadratically with the number of categories, which should therefore be kept low.

state space from (4) is augmented by a discrete state variable a_t.

$$p(x_t, a_t | x_{t-1}, a_{t-1}) \propto \begin{cases} p_{noswitch} & p^{a_t}(x_t|x_{t-1}) \quad \text{if } a_t = a_{t-1} \\ p_{switch} & p^{a_{t-1}\to a_t}(x_t|x_{t-1}) \quad \text{else} \end{cases} \tag{5}$$

Here, the motion model for the case of activity switching $p^{a_{t-1}\to a_t}(x_t|x_{t-1})$ is modelled as a normal distribution around the pose predicted by the regressor $f_{switch}^{a_{t-1}\to a_t}$. The probabilities that an activity transition does or does not occur are denoted p_{switch} and $p_{noswitch}$. In the case of more than two activity categories, these transition probabilities could be represented as a transition matrix with the $p^a_{noswitch}$ of the different categories a on the diagonal.

3.2 Appearance Model

The representation of the subject's image appearance is based on a rough figure-ground segmentation. Under realistic imaging conditions, it is not possible to get a clean silhouette, therefore the image descriptor has to be robust to noisy segmentations to a certain degree. We consider two types of image descriptors, *distance transforms* $dt(Y)$ (Bailey 2004) of segmented figures with a subsequent linear PCA dimensionality reduction step (see Fig. 2c), and a representation obtained by applying *Binary PCA (BPCA)* (Zivkovic and Verbeek 2006) to binary foreground images. Both image descriptors are computed from the content of a bounding box around the centroid of the figure, and 10 to 20 PCA resp. BPCA components have been found to yield good reconstructions. We introduce the following notation for the computation of these descriptors and the projection on the respective subspaces given the raw pixel image Y:

$$\begin{aligned} y_{DT} &= V(dt(Y) - \mu) \\ y_{BPCA} &= BPCA(Y) \end{aligned} \tag{6}$$

In this equation, μ and V are the mean and basis vectors obtained by PCA. $BPCA(Y)$ and $dt(Y)$ are nonlinear operations, in the BPCA case the projection on the subspace is done iteratively (see Zivkovic and Verbeek 2006). As we will see later, it is useful in some situation to consider the inverse operation that projects the image descriptors y_{DT} and y_{BPCA} back into high dimensional pixel space and transforms it into binary images or foreground (fg) probability maps. From the descriptors we compute probability maps via the sigmoid function $\sigma(.)$. In the case of the distance transformed descriptor this is based on the intuition that the foreground/background probabilities are higher far away from the silhouette, and lower very close to the silhouette. BPCA reconstruction is also based on the sigmoid function (Zivkovic and Verbeek 2006).

$$\begin{aligned} p(Y = fg|y_{DT}) &\propto \sigma(V^T y_{DT} + \mu) \\ p(Y = fg|y_{BPCA}) &\propto \sigma(V^T y_{BPCA} + \mu) \end{aligned} \tag{7}$$

Again, μ and V are the mean and basis vectors from linear resp. binary PCA.

Now that we have seen how to compute image descriptors from segmented images and back, we will look how the image appearance is linked to the LLE body pose representation x. We will model the *generative* mapping from pose x to image descriptors y that allows to predict image appearance given pose hypotheses and fits well into generative inference algorithms such as recursive Bayesian sampling. In addition to the local body pose x, the appearance depends on the global body orientation ω (rotation around vertical axis).

$$\begin{aligned} p(y|x,\omega) &= \mathcal{N}(y; f_a(x,\omega), \Sigma_a) \\ f_a(x,\omega) &= W_a \Phi_a(x,\omega) \end{aligned} \tag{8}$$

Here, the functional mapping $f_a(x,\omega)$ is approximated by a sparse kernel regressor (RVM) with weight matrix W_a and kernel functions $\Phi_a(x)$. Σ_a is the prediction variance matrix, it indicates which dimensions of the descriptor y can be well predicted and which cannot, and thus accounts for the fact that the prediction of y will always be subject to uncertainty. Σ_a is estimated from a hold-out set of the original training data and restricted to a diagonal matrix for simplicity.

4 Inferring Image Position, Orientation, Activity and Pose

In this section we will show how the 2D image position, body orientation, activity category, and body pose of the subject are simultaneously estimated given a video sequence, by using the learned models from the previous section within the framework of recursive Bayesian sampling. Both pose estimation and image localisation can benefit from the coupling of pose and image location. For example, the known current pose and motion pattern can help to track through occlusions and distinguish subjects from each other. We therefore believe that tracking should happen jointly in the entire state space Θ,

$$\Theta_t = [a_t, \omega_t, u_t, v_t, w_t, h_t, x_t], \tag{9}$$

consisting of the discrete activity a, orientation ω, the 2D bounding box parameters (position, width and height) u, v, w, h, and the body pose x.

Despite the reduced number of pose dimensions, we face an inference problem in 10-dimensional space. Having a

good sample proposal mechanism like our dynamical model is crucial for the Bayesian recursive sampling to run efficiently with a moderate number of samples. For the monocular sequences we consider, the posteriors can be highly multimodal. For instance a typical walking sequence, *e.g.* observed from a side view, has two obvious posterior modes, shifted 180 degrees in phase, corresponding to the left resp. the right leg swinging forward. When taking the orientation of the figure into account, the situation gets even worse, and the modes are no longer well separated in state space, but can be close in both pose and orientation. Our experiments have shown that a strong dynamical model is necessary to avoid confusion between these posterior modes and reduce ambiguities. Some posterior multimodalities do however remain, since they correspond to a small number of different interpretations of the images, which are all valid and feasible motion patterns.

The precise inference algorithm is very similar to classical *CONDENSATION* (Isard and Blake 1998a), with normalisation of the weights and resampling at each time step. If we neglect the activity switching mechanism for a moment, the prior and likelihood for our inference problem are obtained by extending (4) and (8) to the full state space Θ. In our implementation, the *dynamical* prior $p_d(\Theta_t^i|\Theta_{t-1}^i)$ serves as the sample proposal function. It consists of the learned dynamical pose prior from (2), and a simple motion model for the remaining state variables $\theta = [\omega_t, u_t, v_t, w_t, h_t]$.

$$p_d(\Theta_t^i|\Theta_{t-1}^i) = p_d(x_t^i|x_{t-1}^i)\mathcal{N}(\theta_t^i; \theta_{t-1}^i, \Sigma_\theta) \tag{10}$$

In practice, one may want to use a standard autoregressive model for propagating θ, omitted here for notational simplicity. We assume statistical independence between the body pose x and the state variables θ in (10), since modelling these dependencies would imply restricted camera motions (*e.g.* static camera). The *static* prior over likely body poses (3) and the likelihood (8) are then used for assigning weights w^i to the samples.

$$w_t^i \propto p(y_t^i|\Theta_t^i)p_s(\Theta_t^i)^{\frac{1}{\lambda}} = p(y_t^i|x_t^i, \omega_t^i)p_s(x_t^i)^{\frac{1}{\lambda}} \tag{11}$$

Here, i is the sample index, and y_t^i is the image descriptor computed at time t from the sampled bounding box (u_t^i, v_t^i, w_t^i, h_t^i). Note that our choice for sample proposal and weighting functions differs from *CONDENSATION* in that we only use one component (p_d) of the prior (4) as a proposal function, whereas the other component (p_s) is incorporated in the weighting function.

4.1 Likelihood Computation in Image Space or on a PCA Subspace

Our framework has a generative flavour, since we model the pdf of the appearance given the body pose in a top-down manner. The computation of the image descriptor and projection on the subspace and back can be issued in both directions, as seen in (6) and (7). One possibility is to compute the image descriptors in a bottom-up manner and project them onto the PCA or BPCA subspace (6), where the likelihood is then directly obtained using (8).

Alternatively, in a purely generative top-down manner, we can predict whether we expect a certain pixel to be foreground or background given a pose hypothesis. This is done by concatenating the mapping $f_a(x, \omega)$ from (8) and the projection of the appearance descriptor into full appearance space (image space) (7). This yields a discrete 2D probability distribution of foreground probabilities *Seg* over the pixels $\mathbf{p}$ in the bounding box. From this pdf, a likelihood measure can then be derived by comparing it to the actually observed segmented image *Obs*, also viewed as a discrete pdf, using the Bhattacharyya similarity measure (Bhattacharyya 1943) which measures the affinity between distributions.

$$\begin{aligned} Seg_t^i(\mathbf{p}) &= p(\mathbf{p} = fg|f_a(x_t^i, \omega_t^i)) \\ Obs_t^i(\mathbf{p}) &= p(\mathbf{p} = fg|image_t, u_t^i, v_t^i, w_t^i, h_t^i) \\ BC_t^i &= \sum_{\mathbf{p}} \sqrt{Seg_t^i(\mathbf{p})Obs_t^i(\mathbf{p})} \end{aligned} \tag{12}$$

Both alternative ways of likelihood computation have advantages and drawbacks. The bottom-up variant requires binary images to compute the image descriptors, whereas the top-down variant can handle continuous foreground probabilities. Often the foreground segmentation is available in the form of probability maps, and thresholding it may cause an unnecessary loss of information and yield unsatisfying results. On the other hand, evaluation of likelihood on the (B)PCA subspace can benefit from the learned variance Σ_a of the appearance prediction. Also, the bottom-up computation of descriptors can be disturbed by noisy segmentations. This holds particularly for the *distance transformed* image descriptor y_{DT}. In the case of the descriptor based on BPCA, the projection on the subspace is iterative and therefore slow, which in this case reduces the attractivity of the bottom-up variant from a practical perspective. Experimentally, the combination of *distance transformed* descriptors and bottom-up descriptor computation fails when the input image segmentation is very noisy, the other three combinations perform similarly well.

4.2 Activity Switching

When turning to the multi activity tracking case, the sample proposal function is adapted according to (5). A sample i undergoes an activity switch with probability p_{switch}. In our experiments, the scheme is demonstrated for two activity categories, *walking* and *running*, therefore we set $p_{switch}^{w\to r} =$

$p^{r\to w}_{switch} = 1 - p_{noswitch}$. In case of an activity switch, the sample i is initialised with a value in LLE pose space of the new activity a_t by sampling from the activity transition function $p^{a_{t-1}\to a_t}(x_t|x_{t-1})$. In such a manner, at each time step a number of samples are generated that allow for a smooth transition into the other activity. If these hypotheses are supported by the image data, they will be selected in the subsequent resampling step and take overhand. The percentage of samples of a certain activity category is a measure for the algorithm's belief about the currently observed action. The image support for the hypotheses is given by the observation likelihood, which is always based on the action specific appearance model (f^w_a resp. f^r_a in (8)).

4.3 Globally Optimal Trajectory

The described sample-based tracking algorithm provides a set of N samples with corresponding weights for each frame of the sequence. This representation of the posterior is not suitable for many purposes, even visualisation is difficult. Furthermore, the posteriors are computed on a per-frame level, *i.e.* at time step t we compute $p(\Theta_t|Y_{1:t})$. Often we are interested in a consistent trajectory through the entire image sequence, *i.e.* in the maximum of the posterior $p(\Theta_{1:T}|Y_{1:T})$ over the poses of *all* time steps, given *all* observations. In other words, we are interested in the value for $\Theta_{1:T}$ with maximal probability rather than marginals for each Θ_t.

In our framework this is achieved by a postprocessing algorithm that finds optimal paths through the set of samples. As shown in Doucet et al. (2000b) the MAP estimate of the state sequence is obtained by a Monte-Carlo (forward) filtering stage, followed by a Viterbi algorithm (Forney 1973) that operates on the samples of the particle filter. In the approach proposed here, the Viterbi algorithm is replaced by the max-product algorithm, which is a generalisation of the Viterbi to soft outputs instead of hard decisions (Wiberg 1996; Kschischang et al. 2001). The max-product algorithm is a variant of the standard belief propagation algorithm (or sum-product algorithm). See Kschischang et al. (2001) or Yedidia et al. (2002) for belief propagation algorithms. These algorithms are discrete by nature, *i.e.* each node of the Markov chain (each time step, see also Fig. 7) has a number of discrete states that in our case is equal to the number of samples N of the particle filter tracking algorithm. The algorithm will thus choose one sample per node to form a trajectory through time and state space that best satisfies both observation likelihood and temporal prior. Instead of finding the optimal trajectory for the entire sequence, the algorithm can also be applied to sub-sequences, in a sliding-window fashion. In practice we use the numerically more stable counterpart of max-product, the min-sum algorithm that performs the computations in negative log space instead of probabilities.

In Isard (2003), Sudderth et al. (2003) unifying frameworks have been presented, that generalise belief propagation to continuous state spaces using Monte-Carlo sampling. They perform filtering and smoothing, forward and backward propagation in a single formulation. These methods are however not applicable here, since they are based on the sum-product algorithm and therefore compute per-node marginals. In the two-stage method proposed here, the particle filtering stage provides the discretisation of the state space that is required by the second stage. What might come a bit counterintuitive at first is the fact that this discretisation is non-uniform, different for each node, and in fact reflects the sample proposal distributions of the filtering stage. Having a look at the algorithm, it is however clear that the *max* operation (in contrast to the marginalisation in the sum-product algorithm) is insensitive to the varying density of the sampling, as long as there are sufficient samples in the area of interest.

More formally, the goal is to find a sequence of state variables $\Theta_{1:T}$ that maximises the global function $p(\Theta_{1:T})$, which is factorised into the product of *observation* functions υ that take into account the image information, and *compatibility* functions ψ of temporally adjacent nodes.

$$p(\Theta_{1:T}) = \frac{1}{Z}\prod_{t=2}^{T}\psi(\Theta_t,\Theta_{t-1})\prod_{t=1}^{T}\upsilon(\Theta_t), \tag{13}$$

where Z is a normalisation constant. The equations from the recursive tracking can be reused as the global function uses the same terms. The *observation* functions $\upsilon(\Theta_t)$ are computed according to (11). In fact we can directly reuse the sample weights computed during tracking. The *compatibility* between neighbouring nodes is given by (10). The max-product resp. min-sum algorithm performs inference in this chain graph by propagating local messages between neighbouring nodes. See Fig. 6 for an example of a globally optimised trajectory.

5 Experiments

5.1 Training

The described models were trained on a database of motion sequences from 6 different subjects, walking and running at 3 speeds per activity (2.5, 4.2, 6 resp. 8, 10, 12 km/h). The data was recorded using an optical motion capture system at a frame rate of 60 Hz and subsampled to 30 Hz. The resulting sequences of body poses were normalised for limb lengths and used to animate a realistic computer graphics figure in order to create matching silhouettes for all training poses. The figure was rendered from different view points, located every 10 degrees in a circle around the figure. Due

to this choice of training data, our system currently assumes that the camera is in an approximately horizontal position. The training set consists of 2178 body poses of each activity. All the kernel regressors were trained using the Relevance Vector Machine algorithm (Tipping 2000), with Gaussian Kernels. Different kernel widths were tested and compared using a crossvalidation set consisting of 50% of the training data, in order to avoid overfitting.

5.2 Tracking

We evaluated our tracking algorithm on a number of different sequences. The main goals were to show its ability to deal with noisy sequences with poor foreground segmentation, image sequences of very low resolution, varying viewpoints through the sequence, and switching between activities. The figures in this section show the body poses of the *optimal trajectory* that was computed according to Sect. 4.3, based on the samples from the recursive Bayesian sampling algorithm.

The particle filtering was performed using a set of 500 samples, leading to a computation time of approx. 2–3 seconds per image frame in unoptimised Matlab code. The sample set is initialised in the first frame as follows. Hypotheses for the 2D bounding box locations are either derived from the output of a pedestrian detector that is run on the first image, or from a simple procedure to find connected components in the segmented image. Pose hypotheses x_1^i are difficult to initialise, even manually, since the LLE parametrisation is not easily interpretable. Therefore, we randomly sample from the entire space of feasible poses in the reduced LLE space to generate the initial hypotheses. Thanks to the low-dimensional representation, this works well, and the sample set converges to a low number of clusters after a few time steps, as desired.

The first experiment (Fig. 4) shows tracking on a standard test sequence[2] from (Sidenbladh et al. 2000), where a person walks in a circle. We segmented the images using

[2] http://www.nada.kth.se/~hedvig/data.html

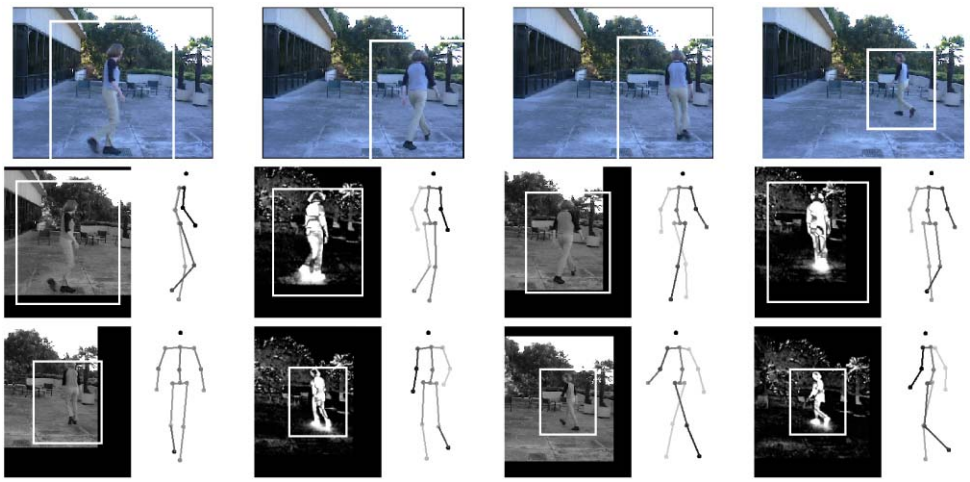

Fig. 4 Circular walking sequence from Sidenbladh et al. (2000). The figure shows full frames (*top*), and cutouts with bounding boxes in original or segmented input images, as well as stick figures of the estimated body poses. For the visualisation of the 3D stick figures, body limbs that are closer in depth appear darker in the plot

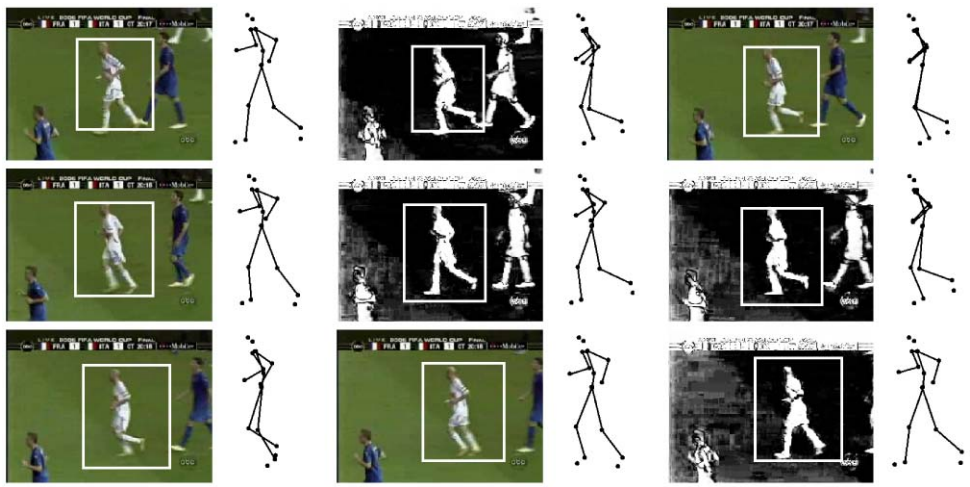

Fig. 5 An extract from a soccer game. The figure shows original and segmented images and with estimated bounding boxes, and estimated 3D poses

background subtraction, yielding noisy foreground probability maps. The main challenge here is the varying viewing angle that is difficult to estimate from the noisy silhouettes. Figure 8 shows another publicly available sequence.[3] Here we used only one camera, while this sequence has been mainly used for multi-camera tracking (*e.g.* Sigal et al. 2004; Sun et al. 2006).

Figure 5 shows an extract from a real soccer game with a running player. The sequence was obtained from www.youtube.com, therefore the resolution is low and the quality suffers from compression artefacts. We obtained a foreground segmentation by masking the colour of the grass.

Figure 9 shows an extract from a treadmill sequence that was 1660 frames long in total. In this sequence, the subject initially walks and switches to running and back to walking several times. The figure shows a few frames from the transition from running to walking; the first two frames clearly contain running poses, then the arms are lowered and the last 3 frames show walking. The plot in Fig. 11a shows the estimated running probabilities throughout this sequence. Even for humans, it is not obvious to identify the exact moment of activity change, there is typically a transition phase of about 0.5 seconds. In our experiments, the activity switch was always detected within this transition phase, as desired. Note that we do not take into account the typical periodic motion in vertical direction that distinguishes running from walking, the activity is correctly estimated from the local shape and its deformation over time alone.

The sequences of Figs. 10 and 12 were recorded in a real traffic environment with a webcam. The image resolution is 320 × 200 pixels, with subjects as small as 40–50 pixels in height. Furthermore, the image quality is unfavourable due to severe MPEG compression artefacts and noisy foreground segmentation that was carried out by subtracting one of the frames at the beginning of the sequence. In Fig. 10 the person carries an umbrella that could be misinterpreted as a leg, and a bag that distorts the overall shape of the pedestrian. The subject also turns away from the camera over the duration of the sequence. Our experiments showed that such a challenging sequence, combining different kinds of difficulties, can only be tracked thanks to the dynamical model, since the information from individual images is unreliable and therefore has to be accumulated over time. The pedestrian in Fig. 12 suddenly starts to run when crossing a street. The activity switch is reliably detected, as can be seen in the activity plot in Fig. 11b.

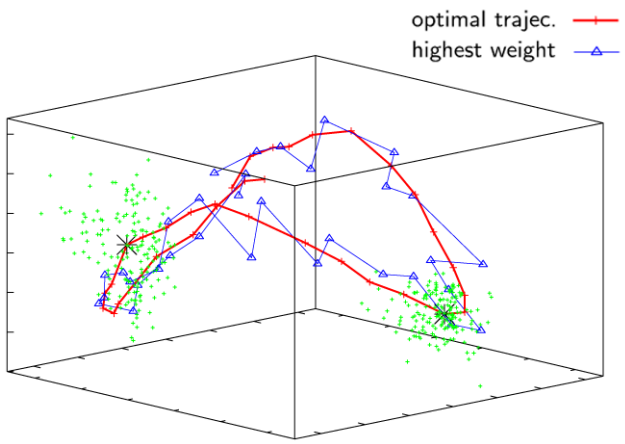

Fig. 6 (Color online) Final trajectory through the LLE pose space obtained by the global optimisation step (*red curve* in the figure). A subsequence of 36 frames, roughly one walking cycle, is shown here. The *blue circles* correspond to the particles with the highest weight, for each timestep of the online tracking algorithm. The *green dots* indicate the sample distribution at frame 4 and 24 of this subsequence

[3] http://www.cs.brown.edu/~ls/

6 Discussion

The proposed approach relies on strong models of prior knowledge about typical human motion patterns. This suggests its use for image sequences, where this prior knowledge is actually needed. For high-resolution multi-camera input sequences, an approach that predominantly relies on the image information might yield more accurate results, and generalise better to unseen motion patterns.

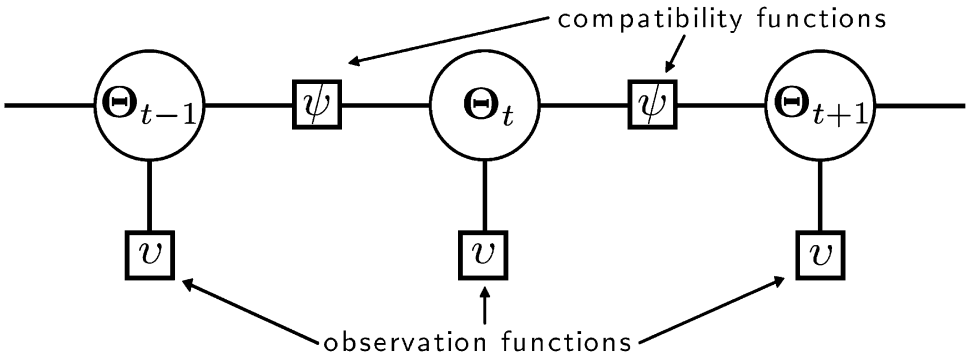

Fig. 7 Graphical model of the Markov chain in which the global optimisation is performed

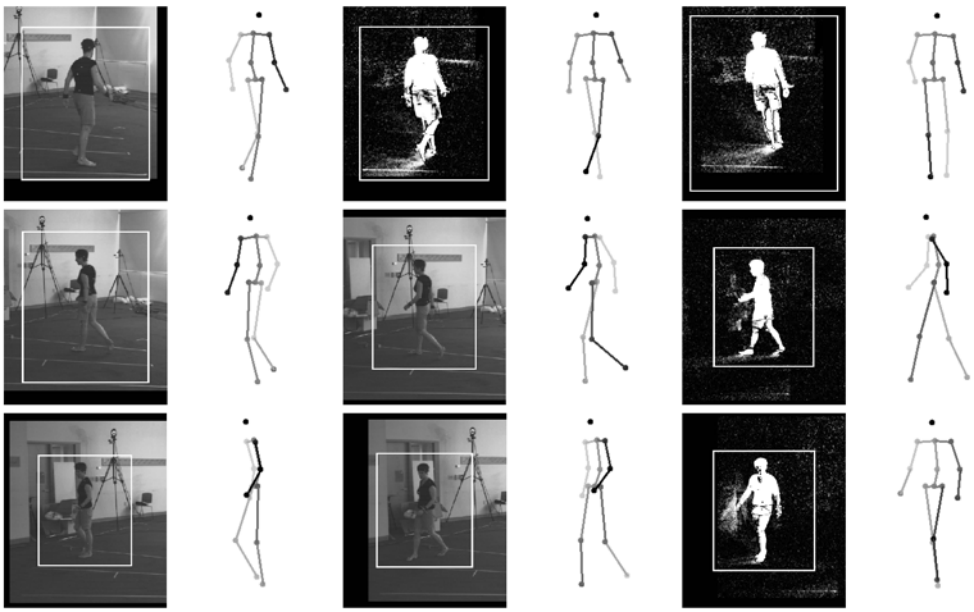

Fig. 8 Circular walking sequence from Sigal et al. (2004), original resp. segmented input images with estimated bounding boxes, and estimated poses

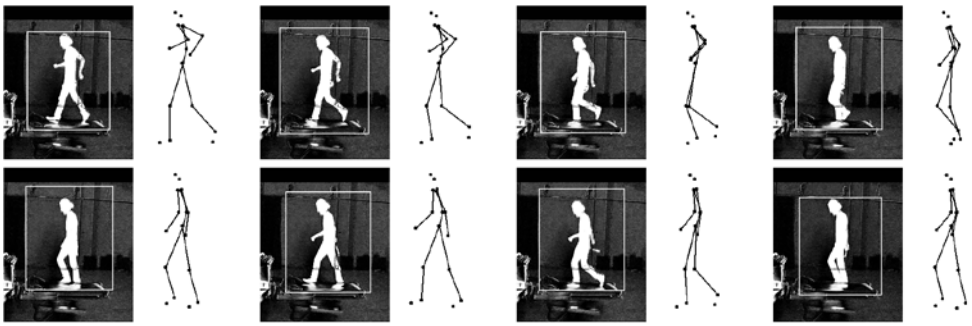

Fig. 9 Transition from running to walking. The original sequence is 1660 frames long, here we show selected frames from the transition phase between frame 921 and frame 936. See also Fig. 11a for a plot of the estimated activity categories of this sequence

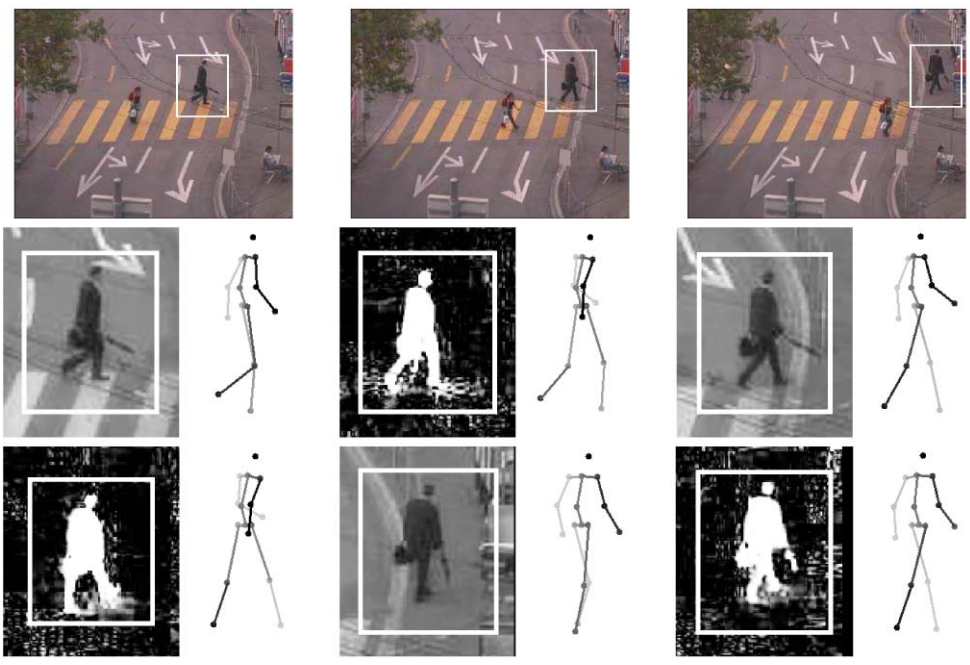

Fig. 10 Real traffic scene with low resolution input images, noisy segmentation, disturbing objects (umbrella, bag), and varying viewangle. Original frames (*top*) and cutouts

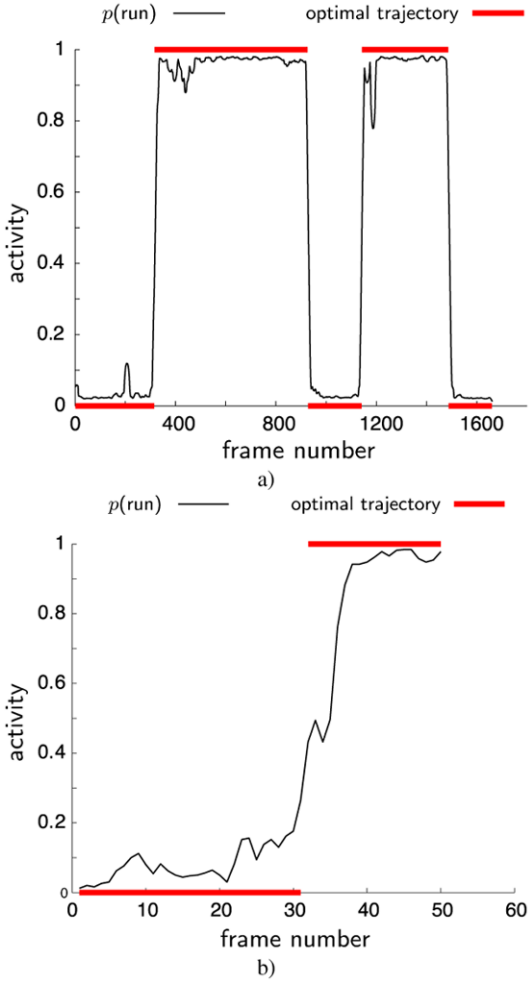

Fig. 11 Activity plots of the sequences of Figs. 9a and 12b. The figures show the estimated activities; the *curve* shows the continuous probability that we observe running rather than walking over the entire sequence, the *bars* indicate the activity label that has been inferred by the global optimisation

The main reasons for failures of the tracking algorithm are excessive noise in the segmented images, especially if the false segmentations are due to occlusions or background objects and thus not randomly distributed. Furthermore, it is very difficult to estimate body poses if the walking direction and view direction of the camera coincide. In such front-views, the image variation that is caused by the body motion is very small, typically much smaller than the image noise, and does thus not allow for successful tracking.

The presented system exhibits complex interactions between its different modules. It would be desirable to evaluate them individually, which is however difficult, because they rely on each other and are often only applicable in combination. For instance, the proposed sampling-based inference scheme requires a low-dimensional pose representation to operate with a moderate number of samples, and so does the learning stage to ensure good generalisation and avoid overfitting. One module that can easily be switched on or off without altering the overall approach is the dynamical model. Here, we have observed that the challenging traffic sequences of Figs. 10 and 12 clearly fail if a simple Brownian motion dynamic model is used instead of the learned model. If the images are of low quality and lack detailed shape information, the scissor-like opening and closing of the legs of a walking person might *e.g.* as well be explained by a backwards walking motion. A low temporal resolution increases the risk of confusion between posterior modes, that can be limited by the dynamical model.

While the activity transition is in general accurately detected, the applied transition model is currently very simple. As there are no activity transitions in the training corpus, the transition itself is not learned. Instead, the transition behaviour is modelled by incorporating the obvious assumption of smooth motion across the activity change, as shown in Sect. 3.1. The results show that the algorithm is able to reliably detect an activity switch and to temporally locate it precisely. Furthermore, the tracked body motion shows a smooth transition from one activity into the other and looks natural. As a possible extension of the system, the actual transition phase could be modelled more accurately by learning from training data as well, including additional body postures that are neither walking nor running poses but occur only during the transition phase.

7 Summary and Conclusion

We presented a monocular tracking approach that simultaneously estimates the 2D bounding box coordinates, the performed activity, and the 3D body pose of a moving person. To this end, we learn statistical models of pose, dynamics, activity transition, and appearance using efficient sparse kernel regressors. The relationship of pose and appearance is learned in a generative manner. Using LLE, we find an embedding of the pose manifolds of low dimensionality, which allows us to use a Monte-Carlo sampling algorithm for tracking. A max-product algorithm finally extracts the optimal sequence through the entire image sequence. We demonstrated the method on several challenging video sequences of low resolution with noisy segmentation.

The activity recognition results reported in this article were nearly perfect, suggesting that the discrimination between the considered activities is a relatively easy task, provided that the tracking works well. Currently, we are applying the proposed method to different data sets with other activity categories than walking and running. This will allow for more detailed conclusions about the potential of our algorithm for the recognition of subtle activity classes. While this article focuses on real-world sequences, a quantitative evaluation of the pose estimation on a benchmark dataset with ground-truth is planned.

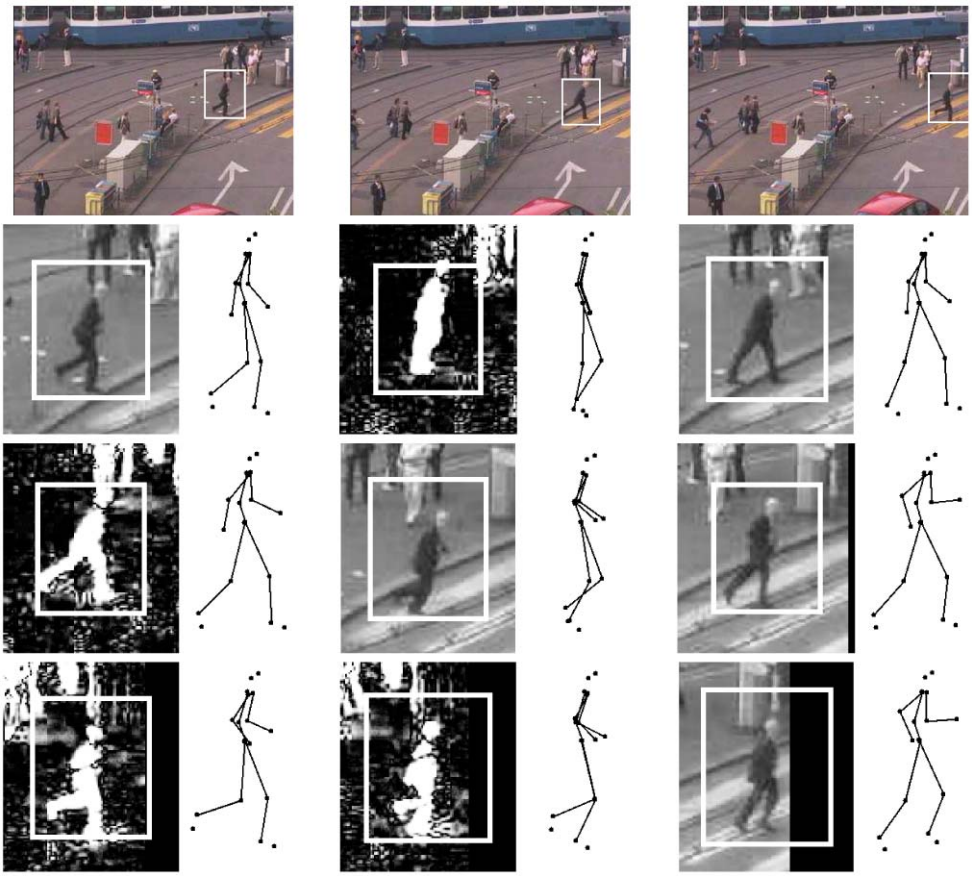

Fig. 12 Real traffic scene with a transition from walking to running. Full frames (*top*) and cutouts with estimated poses. Figure 11b shows the inferred activity categories of this sequence

A further line of current research in tracking related areas is the investigation of other appearance descriptors, and methods to extract interesting features from image data in a statistically meaningful way. One goal is to eventually avoid the need for a segmentation of the images. A different strategy that will be considered is a deeper integration of image segmentation, 2D tracking and 3D pose estimation, where the interaction between these different stages will be investigated.

Acknowledgements This work is supported, in parts, by the FP6 EU Integrated Project DIRAC (IST-027787), the SNF project PICSEL and the SNF NCCR IM2.

References

Agarwal, A., & Triggs, B. (2004a). 3D human pose from silhouettes by relevance vector regression. In *IEEE conference on computer vision and pattern recognition* (*CVPR*).

Agarwal, A., & Triggs, B. (2004b). Tracking articulated motion using a mixture of autoregressive models. In *European conference on computer vision* (*ECCV*).

Agarwal, A., & Triggs, B. (2005). Monocular human motion capture with a mixture of regressors. In *IEEE CVPR workshop on vision for human-computer interaction*.

Bailey, D. G. (2004). An efficient euclidean distance transform. In *International workshop on combinatorial image analysis*.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*.

Doucet, A., Godsill, S., & Andrieu, C. (2000a). On sequentional Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*.

Doucet, A., Godsill, S., & West, M. (2000b). Monte Carlo filtering and smoothing with application to time-varying spectral estimation. In *IEEE conference on acoustics, speech and signal processing* (vol. II, pp. 701–704).

Elgammal, A., & Lee, C.-S. (2004). Inferring 3D body pose from silhouettes using activity manifold learning. In *IEEE conference on computer vision and pattern recognition* (*CVPR*).

Forney, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, *61*(3), 268–278.

Forsyth, D. A., Arikan, O., Ikemoto, L., Brien, J. O., & Ramanan, D. (2006). Computational studies of human motion: Part 1. *Computer Graphics and Vision*, 1(2/3).

Grauman, K., Shakhnarovich, G., & Darrel, T. (2003). Inferring 3D structure with a statistical image-based shape model. *International conference on computer vision* (*ICCV*).

Isard, M. (2003). Pampas: Real-valued graphical models for computer vision. In *IEEE conference on computer vision and pattern recognition* (*CVPR*).

Isard, M., & Blake, A. (1998a). Condensation—conditional density propagation for visual tracking. *International Journal of Computer Vision*, *29*(1), 5–28.

Isard, M., & Blake, A. (1998b). A mixed-state CONDENSATION tracker with automatic model-switching. In *International conference on computer vision* (*ICCV*) (pp. 107–112).

Jaeggli, T., Koller-Meier, E., & Gool, L. V. (2006). Monocular tracking with a mixture of view-dependent learned models. In *IV conference on articulated motion and deformable objects* (*AMDO*).

Kschischang, F., Frey, B. J., & Loeliger, H.-A. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, *47*(2), 498–519.

Lawrence, N. D. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, *6*, 1783–1816.

Lee, C.-S., & Elgammal, A. (2007). Modeling view and posture manifolds for tracking. In *International conference on computer vision* (*ICCV*).

Li, R., Yang, M.-H., Sclaroff, S., & Tian, T.-P. (2006). Monocular tracking of 3D human motion with a coordinated mixture of factor analyzers. In *European conference on computer vision* (*ECCV*) (pp. 137–150).

Li, R., Tian, T.-P., & Sclaroff, S. (2007). Simultaneous learning of nonlinear manifold and dynamical models for high-dimensional time series. In *International conference on computer vision* (*ICCV*).

Lim, H., Camps, O. I., Sznaier, M., & Morariu, V. I. (2006). Dynamic appearance modeling for human tracking. In *IEEE conference on computer vision and pattern recognition* (*CVPR*) (pp. 751–757).

Moeslund, T., Hilton, A., & Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, *104*(2), 90–126.

Navaratnam, R., Fitzgibbon, A. W., & Cipolla, R. (2007). The joint manifold model for semi-supervised multi-valued regression. In *International conference on computer vision* (*ICCV*).

Pavlovic, V., Rehg, J. M., & MacCormick, J. (2001). Learning switching linear models of human motion. In *Neural information processing systems*.

Rosales, R., & Sclaroff, S. (2001). Learning body pose via specialized maps. In *Neural information processing systems*.

Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326.

Sidenbladh, H., Black, M., & Fleet, D. (2000). Stochastic tracking of 3D human figures using 2D image motion. In *European conference on computer vision* (*ECCV*) (pp. 702–718).

Sigal, L., Bhatia, S., Roth, S., Black, M., & Isard, M. (2004). Tracking loose-limbed people. In *IEEE conference on computer vision and pattern recognition* (*CVPR*).

Sminchisescu, C., & Jepson, A. (2004). Generative modeling for continuous non-linearly embedded visual inference. In *International conference on machine learning* (*ICML*).

Sminchisescu, C., Kanaujia, A., Li, Z., & Metaxas, D. (2005). Discriminative density propagation for 3D human motion estimation. In *IEEE conference on computer vision and pattern recognition* (*CVPR*).

Sudderth, E. B., Ihler, A. T., Freeman, W. T., & Willsky, A. S. (2003). Nonparametric belief propagation. In *IEEE conference on computer vision and pattern recognition* (*CVPR*).

Sun, Y., Bray, M., Thayananthan, A., Yuanand, B., & Torr, P. (2006). Regression-based human motion capture from voxel data. In *British machine vision conference*.

Tenenbaum, J., de Silva, V., & Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

Thayananthan, A., Navaratnam, R., Stenger, B., Torr, P., & Cipolla, R. (2006). Multivariate relevance vector machines for tracking. In *European conference on computer vision* (*ECCV*).

Tipping, M. (2000). The relevance vector machine. In *Neural information processing systems*.

Urtasun, R., Fleet, D. J., & Fua, P. (2006). 3D people tracking with Gaussian process dynamical models. In *IEEE conference on computer vision and pattern recognition* (*CVPR*) (pp. 238–245).

Wang, J. M., Fleet, D. J., & Hertzmann, A. (2006). Gaussian process dynamical models. In *Neural information processing systems* (pp. 1441–1448).

Wiberg, N. (1996). *Codes and decoding on general graphs*. PhD thesis, Department of Electrical Engineering, Linköping University, Sweden.

Yedidia, J., Freeman, W., & Weiss, Y. (2002). *Understanding belief propagation and its generalizations* (Technical report TR-2001-22). MERL.

Zivkovic, Z., & Verbeek, J. (2006). Transformation invariant component analysis for binary images. In *IEEE conference on computer vision and pattern recognition* (*CVPR*) (pp. 254–259).