Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST - Priority 2

DELIVERABLE NO: D 3.4
Biological Motion Detection Based on the Learned, Statistical Representations

Date of deliverable: 30.06.2007
Actual submission date: 07.08.2007

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: **ETH Zurich**

Revision [1]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

D3.4 Biological Motion Detection Based on the learned, Statistical representations

ETH Zurich (ETHZ)
Katholieke Universiteit Leuven (KUL)

Abstract:
This deliverable presents two contributions. The first is a statistical analysis of the human motion dataset that was produced in D3.2 with a Motion Capture (MoCap) setup. The purpose of this dataset is to develop and evaluate both biological and computational models for action recognition based on the same underlying data, so that later experimental results become comparable and findings can be transferred. The statistics of the motion data are interesting, since they determine, to a large extent, the core of the machine learning techniques we consider for our computer vision algorithms. In addition, knowledge about the stimuli's underlying dimensionality and principal modes of variation is important in order to design and evaluate neurophysiologic experiments.

The second contribution is a newly developed computational body pose estimation and action recognition algorithm that is trained on the recorded MoCap data. It is specifically designed to handle the uncertainty and ambiguities that are inherent in noisy real-world video input using a robust statistical model. The algorithm uses the learned relationship between body pose and appearance available from the MoCap data both to constrain individual body poses and their sequence in the context of an activity model. Body pose estimation and tracking is achieved by a recursive Bayesian sampling algorithm with an activity switching mechanism based on learned transfer functions. The approach is experimentally evaluated on several challenging video sequences and achieves good results in difficult real-world settings.

Table of Content

1. Introduction

A major line of research in DIRAC WP3 aims at drawing parallels between biological visual perception (in animals) and computational processing/analysis of similar visual stimuli. On the biological side, insights are gained by presenting test subjects (macaque monkeys) with visual stimuli of human actions and analyzing their neural responses. Deliverable D3.3 summarizes the results of a first round of experiments for this research direction.

On the computational side, we mainly operate with statistical tools and statistical machine learning techniques in order to build computational systems that can recognize the same kinds of actions. The motivation behind this research direction is to find computational mechanisms that can successfully deal with the uncertainty, ambiguity, and noise inherent in real-world sensory data. The design of those methods is informed by findings from the biological side, and they can in turn be used to suggest more detailed biological experiments.

The unique approach pursued in DIRAC is to base both kinds of analysis on the same data. For this purpose, we have recorded a dataset of human motions with a Motion Capture (MoCap) setup, as described in deliverable D3.2. The recorded motion data is then used both to create stimuli for neurophysiological experiments on monkeys and to develop and train computational action recognition methods. Although the experimental conditions between both kinds of experiments will differ for practical reasons, the reliance on a common basis will allow us to get a better understanding of the underlying action recognition mechanisms, as well as of the used stimuli and their potentials and limitations.

The purpose of this deliverable is twofold. A first focus lies on the statistical analysis of the dataset of human motions that was produced within the frame of the DIRAC project. The statistics of the motion data are interesting in two respects. On the one hand, they build the core of the machine learning techniques we consider for our computer vision algorithms. On the other hand, they may potentially give insight in principles of biological visual processing. The second focus lies on developing a computational body pose estimation and action recognition algorithm that is trained on the recorded MoCap data and operates on standard video input. This algorithm builds upon the statistical model learned from MoCap data both to constrain individual body poses to physically plausible limb configurations and to constrain their succession in the context of an executed action. As our experiments will show, the developed algorithm achieves successful body pose tracking and action recognition in difficult real-world video sequences.

This document is structured as follows. In Section 2 we will recapitulate the human motion dataset that is the base for the subsequent analysis in Section 3. In Section 4 the computational tracking and activity inference algorithm is summarized. A more detailed technical description is attached to this report. Conclusions are given in Section 5.

2. Human Action Dataset

The basis for the experiments in this deliverable is the Human Action dataset recorded at ETH for deliverable D3.2 (“Setup and Stimuli for Neurophysiological Experiments”) using a MoCap setup. The dataset concentrates on the typical visual motion patterns for two types of human locomotion, *walking* and *running*. Multiple subjects were recorded under laboratory conditions performing those activities at different speeds. The resulting three-dimensional motion data was then further processed, and transferred into representations that are suitable for the planned experiments in the neurophysiological and computational domain. The

following sections briefly recapitulate the recording setup and its use for data acquisition and preprocessing.

2.1 Motion Capture Setup

The sequences were recorded using the ETH Motion Capture setup described in D3.2. This setup consists of an optical MoCap system (VICON) with 6 cameras that operate in the near-infrared range. In order to reconstruct the 3D body motions, 41 infrared-reflective markers are attached to the skin of the test subjects according to a specific protocol. The trajectories of these markers are then tracked in the individual camera streams and integrated into a 3-dimensional representation. Finally, an abstract body model (bi-ped, 17 rigid limbs) is used to interpret that data and solve for body poses. The system operates at 120Hz, and its spatial accuracy is better than 1cm for a working volume of approximately 2m by 2m. In order to capture locomotion patterns of a certain duration, a treadmill was used to keep the recorded person at a stationary location.

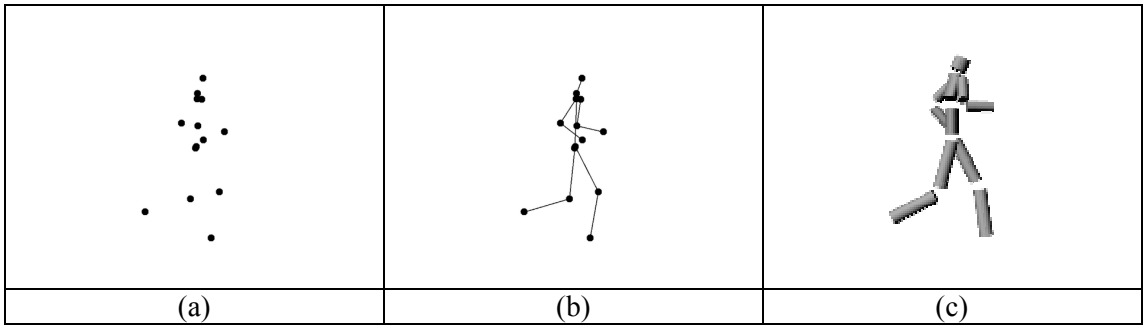
2.2 Human Action Sequence

Six subjects, male, between 20 and 40 years of age, of average physical constitution and in good health, were asked to perform a set of activities on the treadmill. They were allowed to acclimate to moving on the treadmill, which may be a bit cumbersome at first and which may else lead to biased or unnatural motions. The subjects were asked to walk and run at six different speeds for about 10 seconds. The speeds ranged from slow walking (2.5 km/h) over average speed to fast walking (4.2 and 6 km/h). Running was performed at 8, 10 and 12 km/h. The result of this stage are motion capture sequences (36, 6 subjects at 6 speeds), represented either as marker trajectories, or as a kinematic tree with 6 degree-of-freedom (DOF) transformations indicating the relative pose of each limb with respect to its parent limb, or with respect to a global coordinate system. The kinematic fit also provides 3d trajectories of the joint locations such as shoulders and knees, which we will use in our pose representation.

2.3 Use for Computational and Neurophysiological Experiments

The recorded action sequences form a common basis for both computational and neurophysiological experiments. The purpose of starting from joint-angle recordings instead of simple video data is that this representation allows us to exercise more control over the recorded data and transform it into different representations suitable for a large variety of experiments. As described in D3.2, we thus created several different stimuli sets for neurophysiological experiments from the raw MoCap data, where human motion is represented by (a) point-light displays, (b) stick figures, and (c) humanoid figures composed of cylinder-like geometrical primitives (see Figure 1). Those stimuli are currently being used in single-cell recording experiments on macaque monkeys performed by KUL; first results will be described in D3.3.

In addition, we used the same raw data to create realistic-looking video input for training our computational algorithm described in Section 4. This was done by fitting a polygonal “skin model” mesh to the skeleton data and animating it with a computer graphics package, driven by the recorded joint angle data. We thus rendered realistic-looking silhouette sequences from a variety of viewpoints, which were used in order to train the appearance-based body pose estimation algorithm (see Figure 2).



(a) (b) (c)

Figure 1: The three different types of stimuli prepared for neurophysiological experiments: (a) point-light displays where the joint positions are indicated by dots; (b) stick figures, where adjacent joints are connected by a skeleton line; (c) humanoid figures, where body limbs are represented by cylinder-like geometrical primitives.

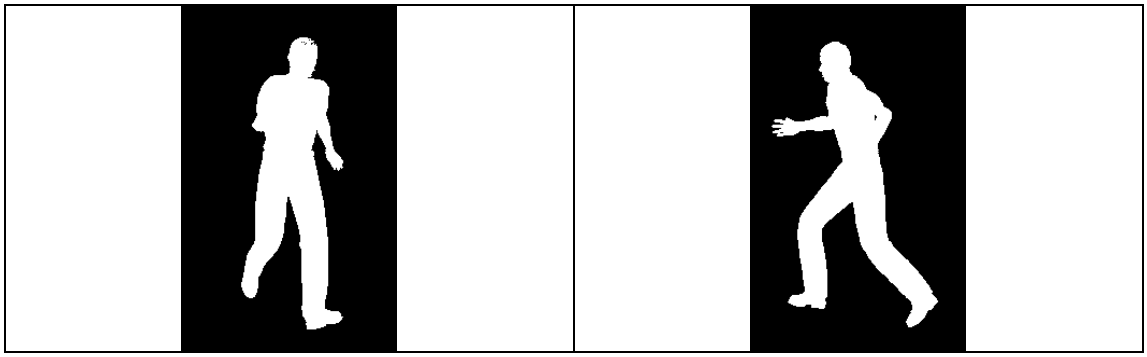


Figure 2: Rendered silhouettes used to train the appearance-based body pose estimation algorithm.

3. Statistical Analysis

In this section, we will have a closer look at the motion data using linear and nonlinear statistical tools. The main purpose is to get a clear idea of the information and variation that is contained in our sequences. This is very important with regard to the interpretation of the neurophysiological experiments. In addition, it directly leads to a lower dimensional compact representation that will be our pose description of choice for the machine learning and inference algorithms.

The analysis targets the intrinsic dimensionality of the data, as well as the main modes of variation of the body pose representation while performing typical movements of human locomotion. Our examination is guided by the proposition that the locomotion trajectories essentially form periodic manifolds of rather low dimensionality that can conceptually be divided into a prototypical motion pattern for the given locomotion type and into modes of variation that characterize the subject specific walking or running styles.

3.1 Data Preparation and Normalization

For the statistical analysis, we use a pose representation that is based on a list of 3D joint locations constituting the overall body configuration. The global translation and rotation of the body is normalized for, since we concentrate on the local body pose for this purpose.
The analysis was carried out on a static snapshot-level, where each body pose from the captured sequences is regarded as one observation. This means that part of the variation in our

dataset stems from the body pose variation within a walking or running cycle, whereas another part stems from the differences of the walking styles between subjects and between activities (running vs. walking).
Alternatively, the data could be analyzed per period, i.e. a sequence of poses corresponding to one walking cycle could be regarded as one observation. This has been done e.g. in [Sidenbladh00, Urtasun04, Jaeggli05] and requires a temporal alignment of the data. Such an analysis would allow for a separation of intra-subject variation within a walking cycle, and inter-subject resp. inter-activity variation.

In order to facilitate the analysis, the data was normalized for limb length, *i.e.* all the motion sequences are based on a single skeleton template. This allows us to concentrate on the body motion and body configuration rather than on anatomical characteristics of the individual test subjects. We thus end up with a 60-dimensional representation that consists of 3d locations such as the head, shoulders, elbows, wrists, and hips.

3.2 Eigenmode Analysis

We first investigated the data using linear tools. The principal components of the 60-dimensional pose examples were extracted. As expected, this analysis showed that our 60-dimensional representation is highly redundant. Indeed, even using a simple linear method such as PCA, 99% of the variation of the data can be explained using only 14 principal components, which corresponds to a reduction of the original dimensionality by more than 75%.

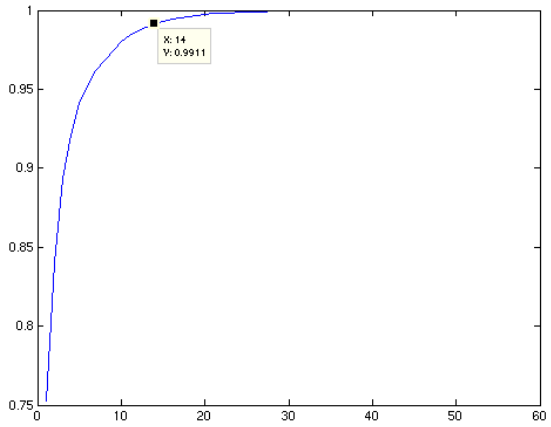
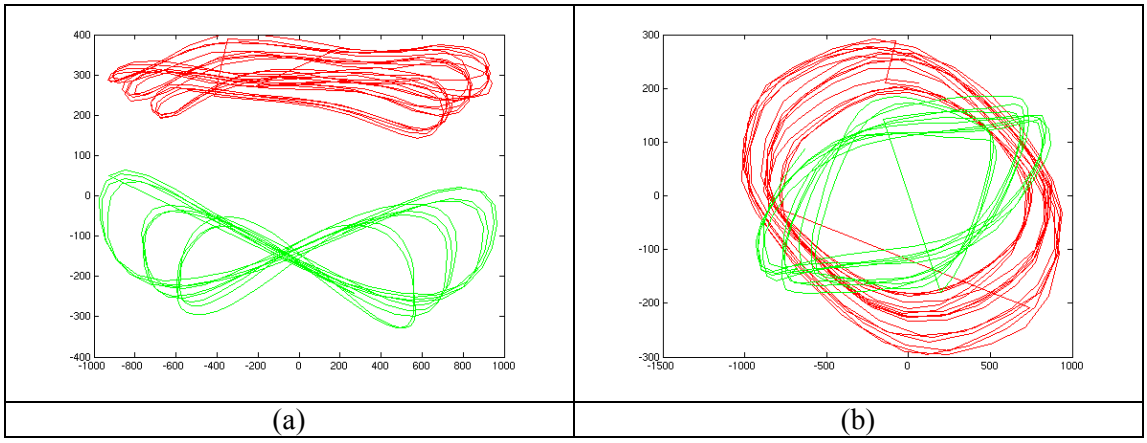


Figure 3: Percentage of the data variance that is captured by a certain number (x-axis) of principal components. The 99% level is reached at 14 principal components.

Another interesting question is whether the motion patterns of the different subjects occupy distinct spaces in the PCA-reduced pose space, e.g. if there is a linear separation that allows for distinguishing between subject or activities given new observations of body poses. In the following, we give a brief overview over the first principal components, aiming at interpreting them in a manner that is meaningful and intuitive for human observers.



(a) (b)

Figure 4: a) Walking sequences (green) and running (red) of a single subject at all speeds. We plotted the first principal component (x-axis) against the second (y-axis). The second component separates walking and running motions. b) The figure shows a plot of the first (x-axis) against the third (y-axis) principal components, for a different subject. In both (a) and (b) the three speeds for each motion pattern can easily be distinguished.

PC1: This PC encodes the main forward and backward motion of the legs during a walking or running cycle. Its amplitude corresponds to the step length and thus correlates with the walking and running speeds. See also Figure 5 and Figure 4.

PC2: variation of this PC results in stick figures that are more or less bent forward and whose knees are bent more or less. This component allows for discrimination between running and walking. Especially for the running examples, it is also strongly dependent on the individual locomotion style of a person, whereas in the walking examples this component mainly varies with the phase of the walking cycle for all subjects (Figure 4(a) and Figure 6).

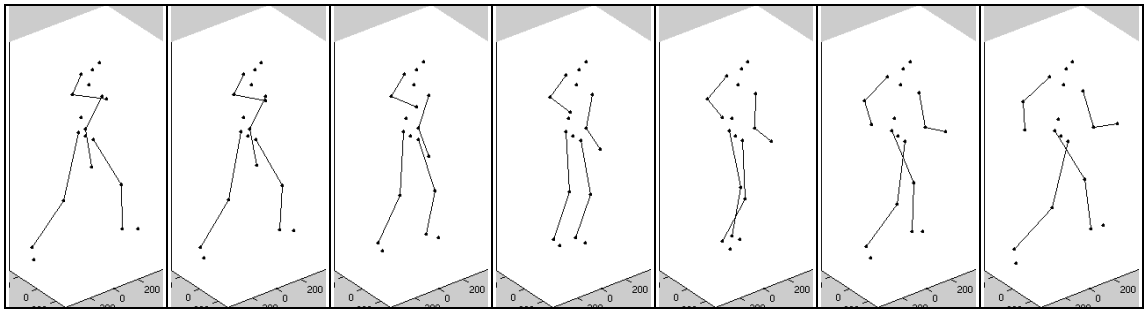


Figure 5: The first principal component captures the main forward/backward motion of the legs. The stick figures were produced by varying the first principal component, while fixing the remaining component at their average value.

PC3: This component periodically varies throughout the walking or running cycle; its amplitude tells how much the free leg is lifted during the floor contact phase of the other leg. This amplitude is consistently higher for running examples than for walking examples; it however depends on the individual running styles and is also positively correlated with the running speed (Figure 7 and Figure 4b).

To summarize, we can state that the first and the third principal components constitute a prototypical walking motion, whereas the remaining components mainly encode subject specific aspects of the locomotion, as well as the transition from walking to running motions.

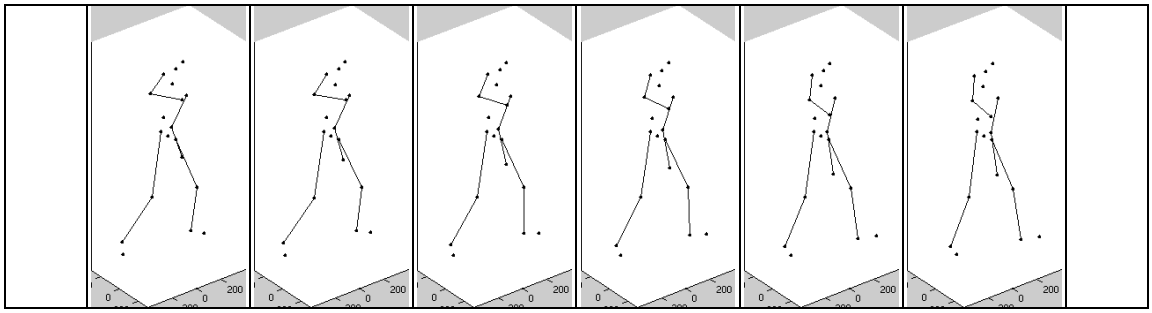


Figure 6: Varying the second principal component.

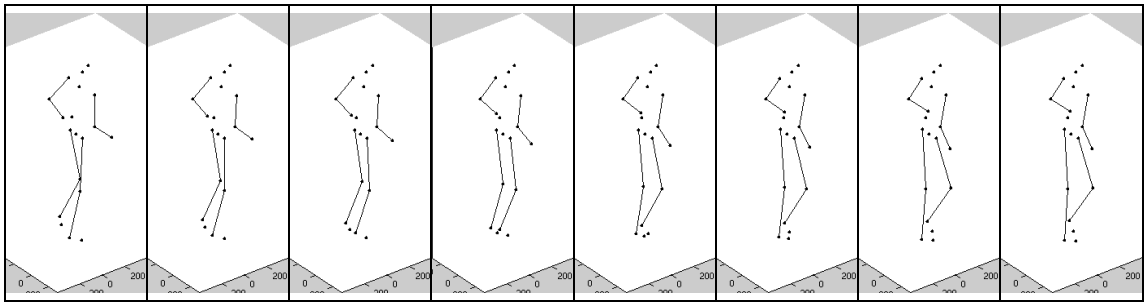


Figure 7: Varying the third principal component results in lifting of the legs.

3.3 LLE Analysis

In the previous section, we have shown that even with linear methods we can obtain a pose representation that is much more compact than the original 60-dimensional one, without losing a prohibitive amount of information. Using non-linear dimensionality reduction methods this can be pushed even further, in order to get an estimate of the true intrinsic dimensionality of the data set. From an engineering point of view, the dimensionality reduction also leads to more compact representations of body poses and thus alleviates the difficulties commonly referred to as the 'curse of dimensionality' in subsequent machine learning and inference stages of our tracking pipeline described in Section 4.

3.3.1 Locally Linear Embedding

LLE was introduced by [Roweis00] as an unsupervised learning algorithm that computes low-dimensional, neighborhood-preserving embeddings of high-dimensional inputs. It maintains the linear local neighborhood relationships of the data points, while allowing for nonlinear deformations of the manifold on a global level.

Unlike other dimensionality reduction techniques, LLE does not provide direct mappings from the high-dimensional to the low-dimensional representation or vice versa. In the frame of our learning based pose estimation algorithm, we need a way to project points on the LLE manifold back into the original pose space. Assuming a functional relationship, we model this mapping with a nonlinear (kernel) regressor.

We computed an LLE embedding for a varying number of LLE dimensions. In its first step, the LLE algorithm describes the relationships of each data point with its nearest neighbors (NNs). The number of NNs taken into account is a parameter that has to be set. In our experiments, we obtained good results with 200 NNs (in a dataset of 2178 data points). Figure 8 shows a 3-dimensional plot of the LLE-reduced walking data, where each test subject has its own color.

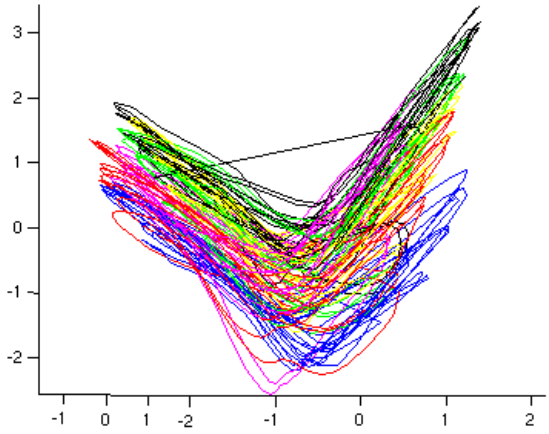


Figure 8: Three-dimensional visualization of the walking motion data using LLE dimensionality reduction.

Figure 9 shows a plot of the reconstruction error as a function of the number of LLE dimensions. This experiment was carried out on all training examples of the class 'WALK'. The reconstruction error is partly due to the regression-based reconstruction mechanism and partly to the loss of information of the actual LLE dimensionality reduction step. The benefit from having more than 6 LLE dimensions in such a setting seems to be rather limited. We have observed that the basic walking motion is well captured by 2-3 LLE components, while additional components allow for the representation of individual walking styles and variation that is due to the different walking speeds.

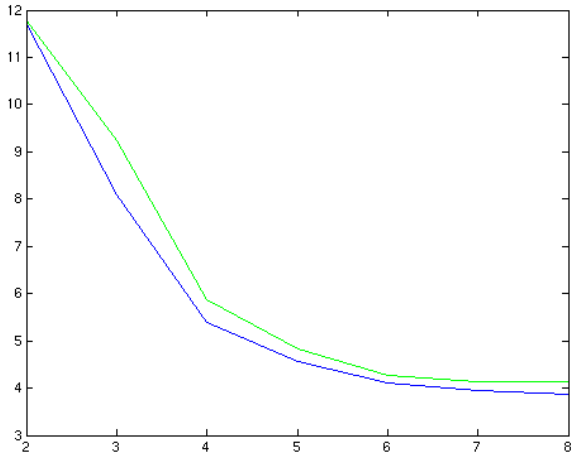


Figure 9: Mean reconstruction errors (in millimeters) as a function of the number of LLE dimensions of the low dimensional body pose representation. The green curve shows results from a hold-out validation set, whereas the blue curve shows the errors of the data that were used to train the kernel regressor for back projection into the original pose space.

3.4 Summary

The investigation suggests that the actual dimensionality is much lower than the 60 dimensions of the original representation. This is essential for the computer vision tracking approach, making inference methods applicable that would be intractable in a high dimensional space. Linear PCA allowed for a dimensionality reduction by more than 75%. Moreover, the first few principal components have a clear interpretation in terms of the prototypical motion pattern or inter-subject vs. inter-activity spreading. Using LLE and a

regression based reconstruction mapping, an activity specific representation of 4-6 dimensions allowed for sufficiently precise projection into the original pose space.

In the following section, the results from this statistical analysis are used in order to develop computational body pose estimation and tracking approach.

4. Tracking Approach

Building upon the recorded MoCap data and using the statistical analysis presented above, ETH has developed a computational algorithm for human body pose estimation and action recognition. This section describes our developed method's core components. For more details, as well as more extensive experimental results, we then refer to the original publications in [ETH-1,ETH-2].

The key idea behind this approach is to model a generative mapping from body poses to appearance descriptors that captures the statistical relationship between individual body poses and their appearance, as well as the dynamics in the pose manifolds. Body poses are inferred by a recursive Bayesian sampling algorithm, which offers a framework for dealing with non-Gaussian and multimodal body pose posteriors.

However, sampling based algorithms are generally not applicable for inference in high-dimensional state spaces, such as the space of body poses. We therefore use Locally Linear Embedding (LLE, [Roweis00]) to find a low-dimensional embedding of our 60-dimensional pose parameterization, using the results from the statistical analysis presented in Section 3. As our results show, the considered motions can be captured reasonably well with just 4 LLE dimensions.

Similarly, we find a low-dimensional embedding for the appearance descriptor that can be compared to the image content both in a top-down and in a bottom-up fashion. Using a rough foreground segmentation, we compare Binary PCA and distance transforms for this modeling step. Sparse kernel regressors are used to capture the non-linearities of the mapping from body pose to appearance efficiently. Although unimodal, the appearance prediction will be subject to uncertainty, since other factors than just the body configuration (pose) may affect appearance (clothing, physical constitution, lighting conditions, etc). This is taken into account by learning a prediction variance matrix of the mapping.

In this work, we consider two action categories: walking and running. Rather than learning a unified representation that contains both walking and running motions, we learn separate activity specific models, as well as a model for transitions between them. The learned transfer functions are used in an activity switching mechanism, which then allows to classify the action that is currently being executed.

The main novelties of this approach are its learned generative appearance modeling (Section 4.2), the tracking in an LLE-reduced pose representation with a nonlinear dynamic model (Sections 4.3-4.5), and the simultaneous recognition of multiple action categories (Section 4.6). A final post-processing step estimates the globally optimal trajectory throughout the entire sequence (Section 4.7). As our experimental results on several challenging sequences show, the dynamical model helps to track through poorly segmented low-resolution image sequences, while at the same time reliably classifying the activity type (Section 4.8 and [ETH-1,ETH-2])

4.1 Tracking Framework

Figure 10 shows an overview of our tracking framework. Its central idea is to express both the high-dimensional body pose description (*i.e.* the joint locations that will be the system's output) and the corresponding image content (*i.e.* the system's input) by lower-dimensional representations, for which it can learn a *generative mapping* from *body pose* to an *appearance descriptor*.

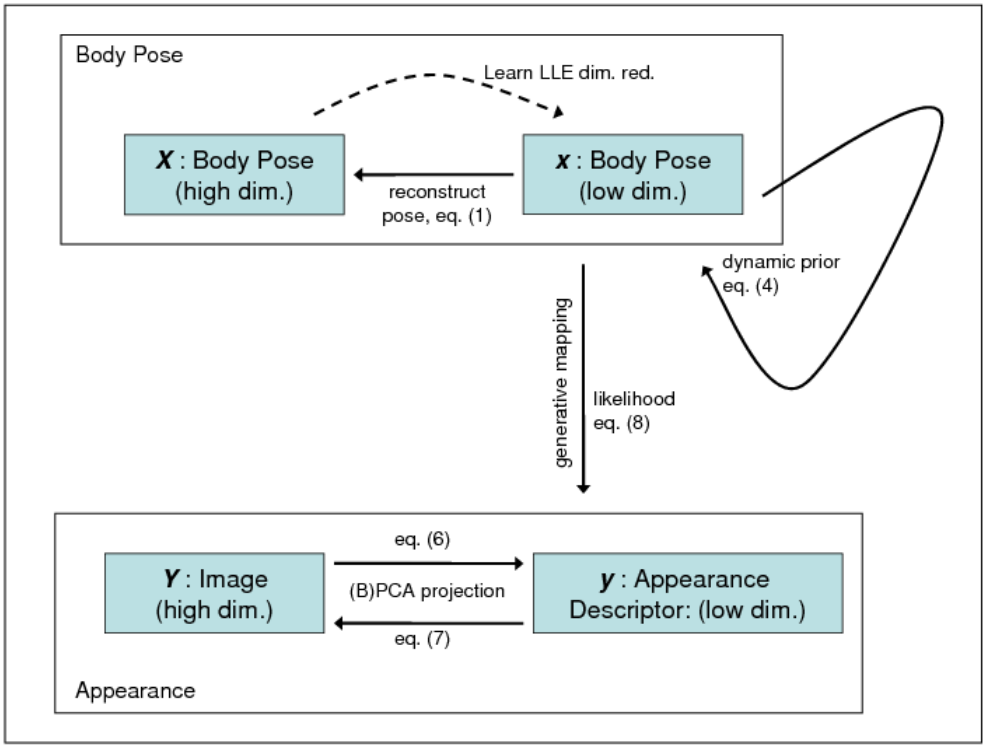


Figure 10: An overview of the general architecture of the tracking framework (the equation numbers refer to the equations in [ETH-1]).

4.2 Appearance Model

The representation of the subject's image appearance is based on a rough figure-ground segmentation. Under realistic imaging conditions, it is typically not possible to get clean silhouettes for a person moving across unconstrained scenes. The image descriptor will therefore have to be to some degree robust to noisy segmentations. We achieve this robustness in two stages.

First, we compute a probabilistic segmentation by modeling the background color by a mixture of Gaussians for each image pixel (similar to [Stauffer99]). Instead of using this model to compute only a hard binary segmentation (Figure 11a), we however also keep the continuous output (see Figure 11b). Secondly, we employ a robust appearance descriptor that models the appearance variability contained in our training data. In [ETH-1], we have experimented with two different choices for this descriptor: a *signed distance transform* [Bailey04] with subsequent linear PCA as a dimensionality reduction step (Figure 11c) and a representation based on *Binary PCA* (BPCA) [Zivkovic06] for binary foreground images. Both image descriptors are computed from the content of a bounding box around the centroid

of the person, and 10-20 PCA resp. BPCA components have been found to yield good reconstructions.

a) b) c)

Figure 11: Binary (a) and continuous (b) foreground segmentation on a real sequence. c) Signed distance transformed silhouette descriptor, computed on clean training silhouette data as used for training the statistical models.

4.3 Pose and Motion Model

Representations for the full body pose are high-dimensional by nature: our current representation is based on 3D joint locations of 20 body locations such as hips, knees, and ankles. To alleviate the difficulties of learning and performing inference in high-dimensional spaces, we identify a low-dimensional embedding of the body pose representation by a dimensionality reduction step. We use LLE [Roweis00], as introduced in Section 3.3.1, which approximately maintains the local neighborhood relationships of each data point, while allowing for global deformations.

LLE dimensionality reduction is performed on all poses in the data set that belong to a certain activity. It expresses each data point in a space of the desired low dimensionality. 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 (RVM). Separate models are learned for the two distinct activities *walking* and *running*.

The training examples form a periodic twisted "ring" in LLE space. As a linear dynamic model is not suitable to predict future poses on this curved manifold, we model the dynamics using another RVM regressor, yielding a *dynamic prior*. This prior is combined with an additional *static prior*, which encodes knowledge about feasible (or likely) body poses for a given activity modeled with a Gaussian Mixture Model (GMM).

Drawing an analogy to the neurophysiological experiments described in D-3.3, this *static prior* corresponds roughly to a *snapshot model* for activity recognition [Giese03]. It only contains information which body poses are likely for a certain activity, but it has no knowledge about their temporal sequence. The *dynamic prior*, on the other hand, encodes such temporal information and describes how the body pose representation may move about in LLE space, but on its own, it may lead to points in this space that do not correspond to valid body configurations. It is important to point out, however, that this dynamic prior does not correspond to the motion pathway postulated in Giese & Poggio's model, which is rather based on low-level motion cues. Instead, it can be seen as a higher-level addition to the form pathway, augmenting it with temporal information.

4.4 Mapping from Pose to Appearance

Next, we model the generative mapping from pose x to image descriptors y that allows to predict image appearance given pose hypotheses and that fits well into generative inference algorithms such as recursive Bayesian sampling. In addition to the local body pose x, the appearance also depends on the viewpoint ω (rotation around the vertical axis) from which the subject is observed. This functional mapping is approximated by a sparse kernel regressor (again an RVM), which is estimated from the training data.

4.5 Interfering Position, Body Pose, Orientation and Activities

Tracking is performed jointly in the entire state space Θ, 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 a 10-dimensional space. Having a good sample proposal mechanism such as our dynamical model is therefore crucial for the Bayesian recursive sampling to run efficiently with a moderate number of samples. The precise inference algorithm is very similar to the classical CONDENSATION [Isard98], but differs in our choice for the sample proposal and weighting functions. The algorithm is described in more detail in [ETH-1].

A special property of our generative framework is that the computation of the image descriptor and its projection onto the subspace and back can be issued in both directions. One possibility is therefore to compute the image descriptors in a bottom-up manner and project them onto the PCA or BPCA subspaces, from which the likelihood can then be directly obtained. 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 back projecting the appearance descriptor into full appearance space and comparing the resulting probability distribution to the observed image using the Bhattacharyya similarity measure [Bhattacharyya43]. Again, please refer to [ETH-1,ETH-2] for details.

Both alternative ways of likelihood computation nicely complement each other. The bottom-up variant requires binary images to compute the image descriptors, while 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 introduce noise. Experimentally, we have found the top-down version to be more robust to such noisy environments. On the other hand, the bottom-up (B)PCA variant can benefit from the learned data covariance matrix, which makes it a good choice in less noisy situations.

4.6 Activity Switching

Each action category has its own low-dimensional pose parameterization expressed in a distinct LLE space. In order to switch between actions, we want to model the transition between those action categories, *i.e.* we want to find walking poses that are very similar to a given running pose and vice versa. During training, we therefore generate two sets of training pairs by looking for the most similar running pose for every walking pose and vice versa. We then model the nonlinear mapping between those pairs using two sparse kernel regressors. This can easily be generalized to more action categories.

During tracking, a number of samples are generated in each time step that allow for a smooth transition into the other activity. If those hypotheses are supported by image information, they will be selected in the subsequent resampling step until they eventually 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.

4.7 Global Trajectory Optimization

The described sample-based tracking algorithm provides a set of N samples with corresponding weights for each frame of the sequence. As we are interested in a consistent trajectory through the entire image sequence, we apply a post-processing algorithm that finds optimal paths through the set of samples. We use *Belief Propagation* with the *Max-Product* algorithm (resp. its numerically more stable *Min-Sum* counterpart) [Kschischang01], which chooses one sample per time step to form a trajectory through time and state space that best satisfies both the observation likelihood and the temporal prior. Instead of searching for the optimal trajectory for the entire sequence, the algorithm can also be applied to sub-sequences in a sliding-window fashion.

4.8 Experimental Results

The tracking algorithm was trained on our human action dataset described in Section 2 using data from all 6 subjects, all 3 speeds per activity, and applying the normalization steps from Section 3.1. In total, the training set consists of 2000 body poses for each activity. All the kernel regressors were trained using the Relevance Vector Machine algorithm with Gaussian kernels [Tipping00]. Different kernel widths were tested and compared using a crossvalidation set consisting of 50% of the training data in order to avoid overfitting.

We then evaluated our tracking algorithm on a number of different test 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.

Particle filtering was performed using a set of 500 samples, leading to a computation time of approximately 2-3 seconds per frame in unoptimized Matlab code. Hypotheses are initialized in the first frame by deriving the subject's position from the output of a pedestrian detector and randomly sampling from the entire space of feasible poses in the reduced LLE representation. This generally works well, and the sample set converges to a low number of clusters within a few time steps, as desired.

Figures 12 and 13 show example results of our algorithm on two challenging sequences (more results can be found in [ETH-1,ETH-2]). The sequence in Figure 12 is a standard test set from [Sidenbladh00], showing a person walking in a circle. The main challenge here is the varying viewing angle that is difficult to estimate from noisy silhouettes. The sequence in Figure 13 was recorded in a real traffic environment with a webcam. It shows a man walking on a pedestrian crossing and then starting to run. The image quality is very low with only 320x240 pixels resolution, subjects as small as 40-50 pixels in height, and severe MPEG compression artifacts. Still, our results show that our approach can successfully process both challenging sequences and additionally detect the transition between the two activities in a reliable fashion.

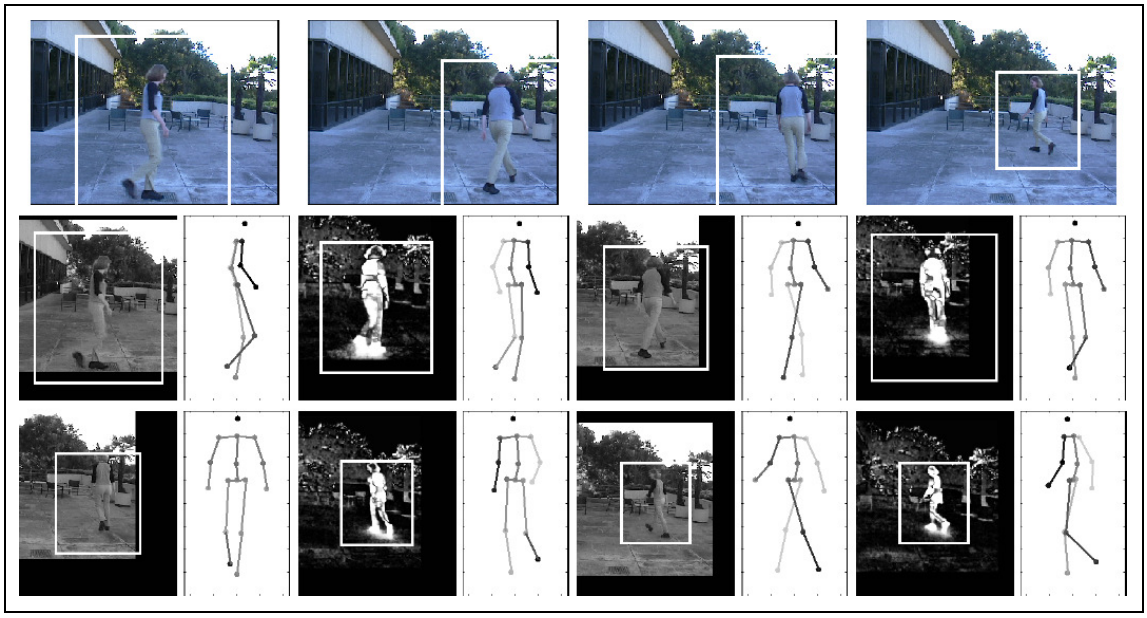


Figure 12: Results for the circular walking sequence from [Sidenbladh00]. The figure shows full frames (top row) and cutouts with bounding box in original or segmented images, together with the estimated poses (middle and bottom row). Darker limbs are closer in depth.

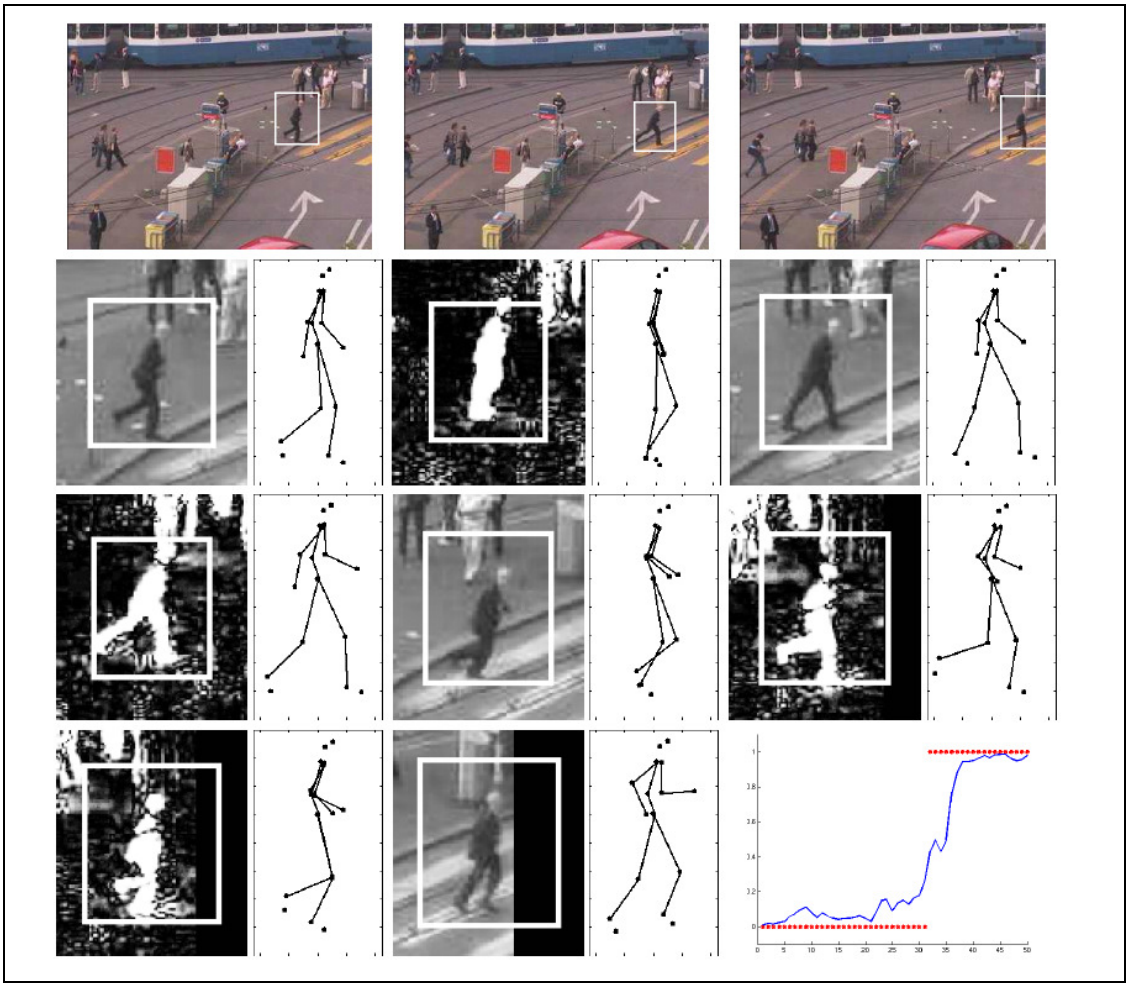


Figure 13: Results for a real traffic scene with a transition from walking to running. The first row shows some examples of the full frames, the other rows contain cutouts with estimated poses. The graph on the bottom right shows the probability of action category *running* (blue solid line) and the activity inferred by the global optimization (red dots).

5. Conclusion

One of the main goals of DIRAC WP3 is to draw parallels between biological and computational action recognition. A concrete strategy pursued in DIRAC towards this goal is to perform experiments and evaluate computational mechanisms for both research fields on the same underlying data. Continuing and complementing work begun in deliverables D3.2 and D3.3, the contribution of this deliverable was therefore twofold. First to perform a statistical analysis of the recorded MoCap human action dataset and thus facilitate neurophysiological and computational experiments. Second to present a computational body pose estimation and action recognition approach based on this data that is able to work with the complexity of real-world video footage.

In future work, the ETH action recognition approach will be gradually extended with different ways to integrate detection and tracking elements in order to reduce pose ambiguity and obtain more reliable results. In particular, we will explore statistical methods to also exploit the correlation between body pose and appearance in order to maximize the discriminative power of the reduced appearance representation. In addition, now that both a biological and a computational setup are available operating on the same data, we will explore cross-links between the two fields and design/perform experiments to cross-check findings from both fields against one another. The results from these experiments will then feed into in the M30 deliverable D3.6.

References

[**ETH-1**] T. Jaeggli, E. Koller-Meier, L. Van Gool. "Multi-Activity Tracking in LLE Body Pose Space", in *ICCV'07 Workshop on Human Motion Understanding, Modeling, Capture, and Animation*, Rio de Janeiro, Brasil, Oct. 2007. **(Appended to this document)**

[**ETH-2**] T. Jaeggli, E. Koller-Meier, L. Van Gool. "Learning Generative Models for Monocular Body Pose Estimation", in *8th Asian Conference on Computer Vision (ACCV'07)*, Tokyo, Japan, Nov. 2007. **(Appended to this document)**

[Bailey04] Bailey, D.G.: An efficient euclidean distance transform. IWCIA (2004)

[Bhattacharyya43] Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math Soc. (1943)

[Giese03] M. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience, 4, 179-192,* 2003.

[Isard98] Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. Int. J. Computer Vision (1998)

[Jaeggli05] T. Jaeggli, G. Caenen, R. Fransens and L. Van Gool, "Analysis of Human Locomotion Based on Partial Measurements", Proceedings of *IEEE Motion* 2005, January 2005

[Kschischang01] Kschischang, F., Frey, B.J., Loeliger, H.A.: Factor graphs and the sum-product algorithm. IEEE Trans. Info. Theory 47, 498–519 (2001)

[Roweis00] S. Roweis, L. Saul. "Nonlinear dimensionality reduction by locally linear embedding." *Science*, v.290 no.5500 , Dec.22, 2000. pp.2323--2326.

[Sidenbladh00] Sidenbladh, H., Black, M., Fleet, D.: Stochastic tracking of 3d human figures using 2d image motion. In ECCV (2000) 702–718

[Stauffer99] C. Stauffer, W.E.L. Grimson, „Adaptive Background Models for Real-Time Tracking“, in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'99)*, 1999.

[Tipping00] Tipping, M.: The relevance vector machine. In: NIPS. (2000)

[Urtasun04] R. Urtasun and P. Fua. Human Motion Models for Characterization and Recognition. In Automated Face and Gesture Recognition, Seoul, May 2004.

[Zivkovic06] Zivkovic, Z., Verbeek, J.: Transformation invariant component analysis for binary images. CVPR (1) 2006: 254-259 (2006)

Learning Generative Models for Monocular Body Pose Estimation

Tobias Jaeggli[1], Esther Koller-Meier[1], and Luc Van Gool[1,2]

[1] ETH Zurich, D-ITET/BIWI, CH-8092 Zurich
[2] Katholieke Universiteit Leuven, ESAT/VISICS, B-3001 Leuven
jaeggli@vision.ee.ethz.ch

Abstract. We consider the problem of monocular 3d body pose tracking from video sequences. This task is inherently ambiguous. We propose to learn a generative model of the relationship of body pose and image appearance using a sparse kernel regressor. Within a particle filtering framework, the potentially multimodal posterior probability distributions can then be inferred. The 2d bounding box location of the person in the image is estimated along with its body pose. Body poses are modelled on a low-dimensional manifold, obtained by LLE dimensionality reduction. In addition to the appearance model, we learn a prior model of likely body poses and a nonlinear dynamical model, making both pose and bounding box estimation more robust. The approach is evaluated on a number of challenging video sequences, showing the ability of the approach to deal with low-resolution images and noise.

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 [1–4]. Generative approaches on the other hand typically use hand crafted geometric body models to predict image appearances (e.g. [5], see [6, 7] for an overview).

We propose to combine the generative methodology with a learning based statistical approach. The mapping from pose to appearance is unimodal and can thus be seen

as a nonlinear regression problem. We approximate the mapping with a RVM kernel regressor [8] that is efficient due to its sparsity.

The human body has many degrees of freedom, leading to high dimensional pose parameterisations. In oder to avoid the difficulties of high dimensionality in both the learning and the inference stage, we apply a nonlinear dimensionality reduction algorithm [9] to a set of motion capture data containing walking and running movements.

1.1 Related Work

Statistical approaches to the monocular pose estimation problem include [1–4, 10, 11]. In [10] the focus lies on the appearance descriptor, and the discriminative mapping from appearance to pose is assumed to be unimodal and thus modelled with a single linear regressor. The multimodality of the discriminative mapping is explicitly addressed in [1–4] by learning *multiple* mappings in parallel as a mixture of regressors. In order to choose between the different hypotheses that the different regressors deliver, [1, 2] use a geometric model that is projected into the image to verify the hypotheses. Inference is performed for each frame independently in [1]. In [2] a temporal model is included using a bank of Kalman filters. In [3, 4] 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 [3], and temporal aspects are included in the learned discriminative mapping, whereas [4] adopts a generative sampling-based tracking algorithm with a first-order autoregressive dynamic model.

These 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 might help the bounding box tracking, and avoid loosing the subject when it is temporarily occluded. To this end, [12] 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 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 [11], where additionally the mapping from the appearance manifold to 3d pose in body joint space is learned using RBF interpolants, allowing for pose inference from sequences of silhouettes.

Instead of modelling manifolds in appearance space, [13–15] work with low dimensional embeddings of body poses. In [13], the low-dimensional pose representation, its dynamics, and the mapping back to the original pose space are learned in a unified framework. These approaches do not include statistical models of image appearance.

In a similar fashion, we also chose to model manifolds in pose space rather than appearance space, because the pose manifold has fewer self-intersections than the appearance manifold, making the dynamics and tracking less ambiguous. In contrast to [13–15], our model includes a learned generative likelihood model. When compared to [1–4, 10, 11], our approach can simultaneously estimate pose and bounding box, and learning a single regressor is more easily manageable than a mixture of regressors.

The paper is structured as follows. Section 2 and 3 introduce our learned models and the inference algorithm, and in Section 4 we show experimental results.

2 Learning

Figure 1 a) shows an overview of the tracking framework. Central element is the low-dimensional body pose parameterisation, 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 real motion capture data sets of

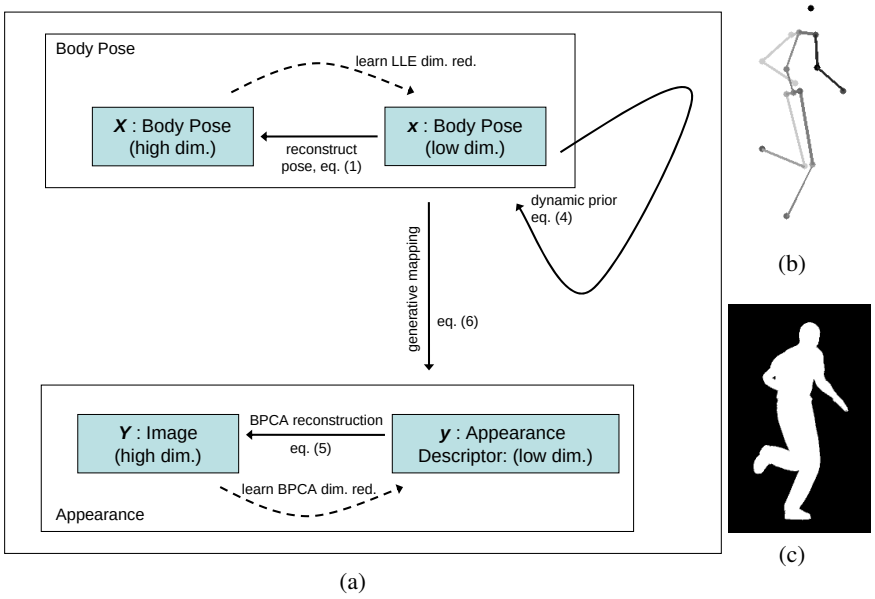


Fig. 1. a) An overview of the tracking framework. Solid arrows represent signal flow during inference, the dashed arrow stands for LLE resp. BPCA dimensionality reduction during training. The figure refers to equations in Section 2. b) Body pose representation as a number of 3d joint locations. c) Corresponding synthetically generated silhouette, as used for training the appearance model.

different subjects, running and walking at different speeds.

2.1 Pose and Motion Prior

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, 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) [9], 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 and expresses each data point in a space of desired low dimensionality. We currently use a 4-dimensional embedding. However, LLE does not provide explicit mappings between the high-dimensional and the low-dimensional space, that would allow to project new data points (that were not contained in the original data set) between the two spaces. 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 (RVM). We use Gaussian kernel functions, computed at the training data locations.

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.

Not all body poses that can be expressed using the LLE pose parameterisation 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. Worse, the learned regressors can produce erroneous outputs when they are applied to unfeasible input poses, since the extrapolation capabilities of kernel regressors to regions without any training data is limited. The additional prior knowledge about feasible body poses 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. We obtain the following formulation for the temporal prior by combination with the *dynamic* prior $p_d(x_t|x_{t-1})$.

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

2.2 Likelihood 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. In order to obtain a compact representation of the appearance of a person, we apply Binary PCA [16] to the binary foreground images. The descriptors are computed from the content of a bounding box around the centroid of the figure, and 10 to 20 BPCA components are kept to yield good reconstructions. The projection of a new bounding box into the BPCA subspace is done in an iterative fashion, as described in [16]. Since we model appearance in a generative top-down fashion, we also consider the inverse operation that projects the low-dimensional image descriptors y back into high dimensional pixel space and transforms it into binary images or foreground probability maps. By linearly projecting y back to the high-dimensional space using the mean μ and basis vectors V of the Binary PCA, we obtain a continuous representation Y_c that is then converted back into a binary image by looking at its signs, or into a foreground probability map via the sigmoid function $\sigma(Y_c)$.

$$p(Y = fg|y) \propto \sigma(V^T y + \mu) \tag{5}$$

Now we will look how the image appearance is linked to the LLE body pose representation x. We model the *generative* mapping f_a 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 particle filtering. In addition to the local body pose x, the appearance depends on the global body orientation ω relative to the camera, around the vertical axis. First, we map the pose x, ω into low dimensional appearance space y,

$$f_a(x, \omega) = W_a \Phi_a(x, \omega) \tag{6}$$

where the functional mapping $f_a(x, \omega)$ is approximated by a sparse kernel regressor (RVM) with weight matrix W_a and kernel functions $\Phi_a(x)$.

By plugging (6) into (5), we obtain a discrete 2d probability distribution of foreground probabilities $Seg(\mathbf{p})$ over the pixels $\mathbf{p}$ in the bounding box.

$$Seg(\mathbf{p}) = p(\mathbf{p} = fg|f_a(x, \omega)) \tag{7}$$

From this pdf, a likelihood measure can then be derived by comparing it to the actually observed segmented image Y_{obs}, also viewed as a discrete pdf $Obs(\mathbf{p})$, using the Bhattacharyya similarity measure [17] which measures the affinity between distributions.

$$\begin{aligned} Obs(\mathbf{p}) =& p(\mathbf{p} = fg|Y_{obs}) \\ BC(x, \omega, Y_{obs}) =& \sum_{\mathbf{p}} \sqrt{Seg(\mathbf{p}) Obs(\mathbf{p})} \end{aligned} \tag{8}$$

We model the likelihood measure as a zero mean Gaussian distribution of the Bhattacharyya distance $d_{Bh} = -ln(BC(x, \omega, Y_{obs}))$, and obtain as the observation likelihood

$$p(Y_{obs}|x, \omega) \propto \exp(-\frac{ln(BC(x, \omega, Y_{obs}))^2}{2\sigma_{BC}^2}) \tag{9}$$

3 Inference

In this section we will show how the 2d image position, body orientation, 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 particle filtering. The pose estimation as well as the image localisation can benefit from the coupling of pose and image location. For example, the known current pose and motion pattern can help to distinguish subjects from each other and track them through occlusions. We therefore believe that tracking should happen jointly in the entire state space Θ,

$$\Theta_t = [\omega_t, u_t, v_t, w_t, h_t, x_t], \tag{10}$$

consisting of the 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 9-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* [18], with normalisation of the weights and resampling at each time step. The prior and likelihood for our inference problem are obtained by extending (4) and (9) 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 prior from eq. (2), and a simple motion model for the remaining state variables $\theta = [\omega_t, u_t, v_t, w_t, h_t]$.

$$p_d(\Theta_t^i|\Theta_{t-1}^i) = p_d(x_t^i|x_{t-1}^i)\mathcal{N}(\theta_t^i; \theta_{t-1}^i, \Sigma_\theta) \tag{11}$$

In practice, one may want to use a standard autoregressive model for propagating θ, omitted here for notational simplicity. The *static* prior over likely body poses (3) and the likelihood (9) are then used for assigning weights w^i to the samples.

$$w_t^i \propto p(Y_t^i|\Theta_t^i)p_s(\Theta_t^i) = p(Y_t^i|x_t^i, \omega_t^i)p_s(x_t^i) \tag{12}$$

Here, i is the sample index, and Y_t^i is the foreground probability map contained in the sampled bounding box $(u_t^i, v_t^i, w_t^i, h_t^i)$ of the actually observed image. 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 Experiments

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 seg-

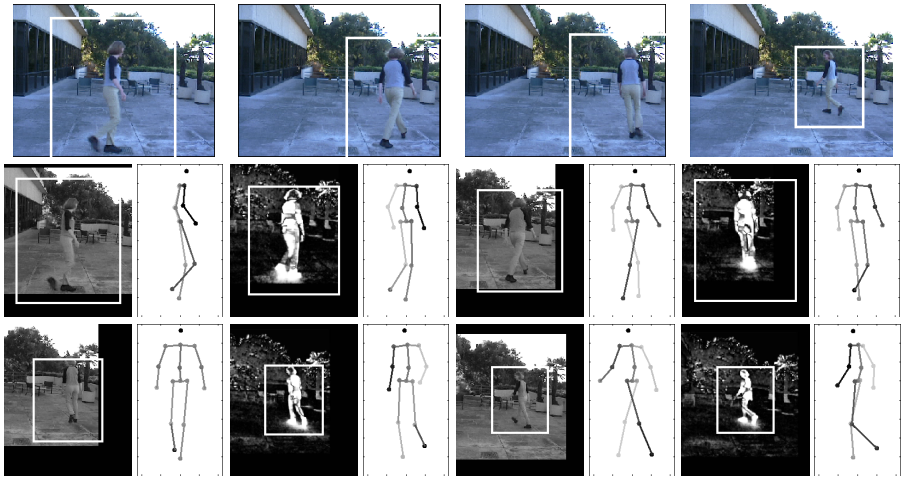


Fig. 2. Circular walking sequence from [5]. The figure shows full frames (top), and cutouts with bounding box in original or segmented input images and estimated poses. Darker limbs are closer in depth.

mentation, image sequences of very low resolution, and varying viewpoints through the sequence.

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 parameterisation 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 described models were trained on a database of motion sequences from 6 different subjects, walking and running at different speeds. The data was recorded using an optical motion capture system. 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 (see Fig. 1c). 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 4000 body poses in total. All the kernel regressors were trained using the Relevance Vector Machine algorithm (RVM) [8], 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. 4 LLE dimensions were used, and 15 BPCA components.

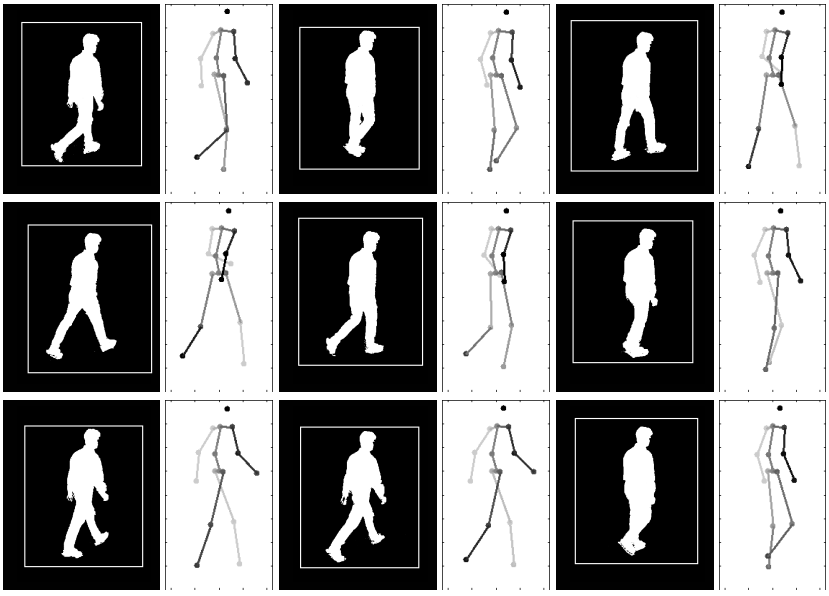


Fig. 3. Diagonal walking sequence. Estimated bounding boxes and poses. The intensity of the stick figure limbs encodes depth; lighter limbs are further away.

The first experiment (Fig. 2) shows tracking on a standard test sequence[3] from [5], where a person walks in a circle. We segmented the images using 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. Tracking through another publicly available sequence from the *HumanID* corpus is shown in Figure 3. The subject walks in an angle of approx. 35 degrees to the camera plane. In addition it is viewed from a slight top-view and shows limb foreshortening due to the perspective projection. These are violations of the assumptions that are inherent in our choice of training data, where we used horizontal views and orthographic projection. Nevertheless the tracker performs well.

Figure 4 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 color of the grass. In Figure 5 we show a real traffic scene that was recorded with a webcam of 320×240 pixels. The subjects are as small as 40 pixels in height. Noisy foreground segmentation was carried out by subtracting one of the frames at the beginning of the sequence.

Our experiments have shown that the dynamical model is crucial for tracking through these sequences, since the image information is unreliable and therefore has to be accumulated over time. The tracker can otherwise be distracted by the noisy segmentations and the multimodal per-frame likelihoods.

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

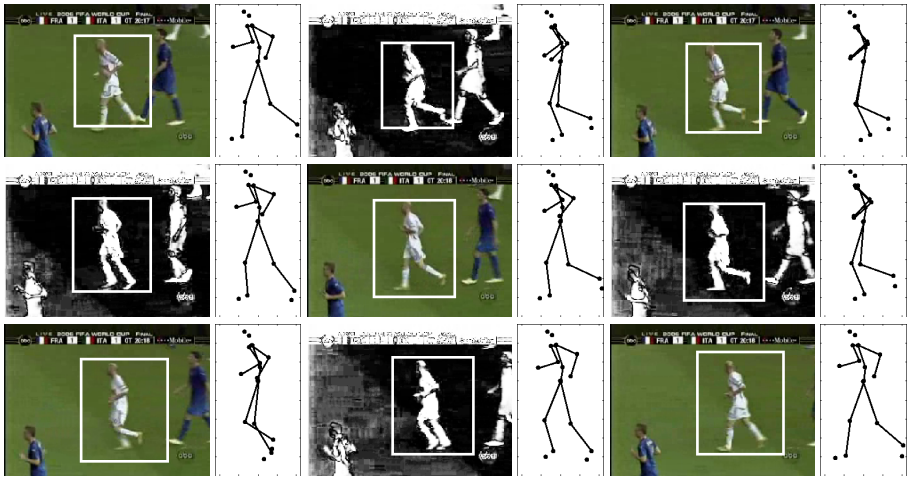


Fig. 4. An extract from a soccer game. The figure shows original and segmented images and with estimated bounding boxes, and estimated 3d poses.

5 Summary and Conclusion

We have proposed a learning-based approach to the estimation of 3d body pose and image bounding boxes from monocular video sequences. The relationship between body pose and image appearance is learned in a generative manner. Inference is performed with a particle filter that samples in a low-dimensional body pose representation obtained by LLE. A nonlinear dynamical model is learned from training data as well. Our experiments show that the proposed approach can track walking and running persons through video sequences of low resolution and unfavourable image quality.

Future work will include several extensions of the current method. We will explicitly consider multiple activity categories and perform action recognition along with the tracking. Also, we will investigate different image descriptors, that do extract the relevant image information more efficiently.

Acknowledgements

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

References

1. Rosales, R., Sclaroff, S.: Learning body pose via specialized maps. NIPS (2001)
2. Thayananthan, A., Navaratnam, R., Stenger, B., Torr, P., Cipolla, R.: Multivariate relevance vector machines for tracking. Ninth European Conference on Computer Vision (2006)
3. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Discriminative density propagation for 3d human motion estimation. CVPR (2005)

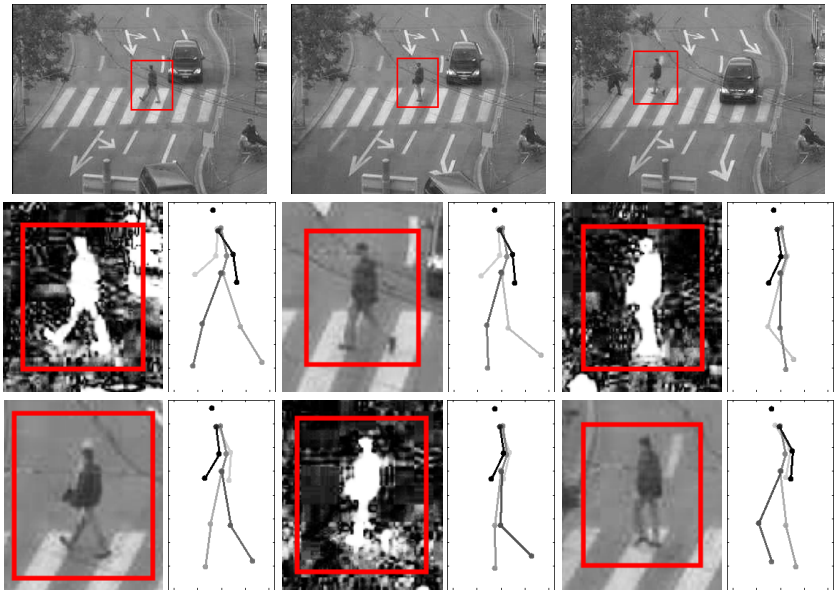


Fig. 5. Traffic scene with low resolution images and noisy segmentation.

4. Agarwal, A., Triggs, B.: Monocular human motion capture with a mixture of regressors. IEEE Workshop on Vision for Human-Computer Interaction at CVPR (2005)
5. Sidenbladh, H., Black, M., Fleet, D.: Stochastic tracking of 3d human figures using 2d image motion. In ECCV (2000) 702–718
6. Forsyth, D.A., Arikan, O., Ikemoto, L., J. O' Brien, D.R.: Computational studies of human motion: Part 1. Computer Graphics and Vision Volume 1 Issue 2/3 (2006)
7. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104**(2) (2006) 90–126
8. Tipping, M.: The relevance vector machine. In: NIPS. (2000)
9. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science, v.290 no.5500 , Dec.22, 2000, pp.2323–2326 (2000)
10. Agarwal, A., Triggs, B.: A local basis representation for estimating human pose from cluttered images. Proceedings of the Asian Conference on Computer Vision (ACCV) (2006)
11. Elgammal, A., Lee, C.S.: Inferring 3d body pose from silhouettes using activity manifold learning. CVPR (2004)
12. Lim, H., Camps, O.I., Sznaier, M., Morariu, V.I.: Dynamic appearance modeling for human tracking. In: Conference on Computer Vision and Pattern Recognition. (2006) 751–757
13. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models. In: Advances in Neural Information Processing Systems 18. (2006) 1441–1448
14. Sminchisescu, C., Jepson, A.: Generative modeling for continuous non-linearly embedded visual inference. International Conference on Machine Learning, ICML (2004)
15. Li, R., Yang, M.H., Sclaroff, S., Tian, T.P.: Monocular tracking of 3d human motion with a coordinated mixture of factor analyzers. ECCV (2) (2006) 137–150
16. Zivkovic, Z., Verbeek, J.: Transformation invariant component analysis for binary images. CVPR (1) 2006: 254-259 (2006)
17. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math Soc. (1943)

18. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. Int. J. Computer Vision (1998)

Multi-Activity Tracking in LLE Body Pose Space

Tobias Jaeggli[1], Esther Koller-Meier[1], and Luc Van Gool[1,2]

[1] ETH Zurich, D-ITET/BIWI, CH-8092 Zurich
[2] Katholieke Universiteit Leuven, ESAT/VISICS, B-3001 Leuven
jaeggli@vision.ee.ethz.ch

Abstract. We present a method to simultaneously estimate 3d body pose and action categories from monocular video sequences. Our approach learns a low-dimensional embedding of the pose manifolds, as well as the statistical relationship between body poses and their image appearance. In addition, the dynamics in these pose manifolds are modelled. Sparse kernel regressors capture the non-linearities of these mappings efficiently. Body poses are inferred by a recursive Bayesian sampling algorithm with an activity-switching mechanism based on learned transfer functions. 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.

1 Introduction

We consider the problem of estimating human body pose and action categories from image sequences. This is a difficult problem, especially when dealing with low quality low resolution imagery. Often the individual images do not provide enough information to resolve ambiguous situations, and strong prior models have to be adopted in order to compensate for that lack of information.

To address these problems we propose a method to estimate 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 as well. All the mappings are modelled using sparse kernel regressors, leading to efficient evaluation during tracking.

The observations are taken into account in the form of roughly segmented images that are obtained by a pre-processing step such as motion segmentation, background subtraction or other. The underlying relationship between image appearance and body poses is multivalued and ambiguous, thus non functional. Other learning based approaches have explicitly modelled the multimodality of the discriminative mapping from appearances to poses (or the joint probability density function between appearance and pose) with mixtures of regressors or Gaussians (e.g. [1–4]). The number of

required regressors is however a delicate parameter of these systems, as is the regularisation during the learning stage, which is needed to avoid overfitting. We therefore follow the opposite strategy, and model the generative mapping from body poses to appearance descriptors, which can be assumed to be functional and thus be approximated with a nonlinear kernel regressor. Although unimodal, 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 a prediction variance matrix of the mapping.

A main focus of the proposed approach lies on the ambiguities and uncertainties that are inherent in monocular body tracking. Recursive Bayesian Sampling 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, [5]) to find a low-dimensional embedding of our 60-dimensional pose parameterisation. With 4 LLE dimensions, the considered motions can be captured reasonably well.

In this paper 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 recognize the performed activity along with the pose estimation, using a switching mechanism of the inference algorithm.

The main novelties of this paper 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.

1.1 Related Work

There is a wide variety of literature about body pose estimation and tracking (see [6] for an overview). Here we will have a look at the application of statistical methods to this problem that infer poses from one or multiple camera streams. Many authors adopt a discriminative strategy to infer poses directly from image descriptors [1–4, 7–9].

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 [9], 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 [8] 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 [1–4] by learning *multiple* mappings in parallel as a mixture of regressors. In order to choose between the different hypotheses that the different regressors deliver, [1, 2] use a geometric model that is projected into the image to verify the hypotheses. Inference is performed for each frame independently in [1]. In [2] 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 [3, 4] 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 [3], and temporal aspects are included in the learned discriminative mapping, whereas [4] adopts a generative sampling-based tracking algorithm with a first-order autoregressive dynamic model.

These 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 might help the bounding box tracking, and avoid loosing the subject when it is temporarily occluded. To this end, [10] 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 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 [11], where additionally the mapping from the appearance manifold to 3d pose in body joint space is learned using RBF interpolants, allowing for pose inference from sequences of silhouettes.

Instead of modelling manifolds in appearance space, [12–14] work with low dimensional embeddings of body poses. In [12], the low-dimensional pose representation, its dynamics, and the mapping back to the original pose space are learned in a unified framework. These approaches do not include statistical models of image appearance. Our approach also models pose manifolds rather than appearance manifolds, because the pose manifold has fewer self-intersections than the appearance manifold, making the dynamics and tracking less ambiguous.

Regarding activity switching, [15] has 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 fundamentally 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 remainder of this paper is organised as follows. Section 2 introduces our learned models. In Section 3 the sample-based inference is presented and Section 4 shows experimental results on different video sequences.

2 Statistical Modelling

Figure 1 a) shows an overview of the tracking framework, reduced to a single activity category for clarity. Central element is the low-dimensional body pose parameterisation, 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 real motion capture data sets of different subjects, running and walking

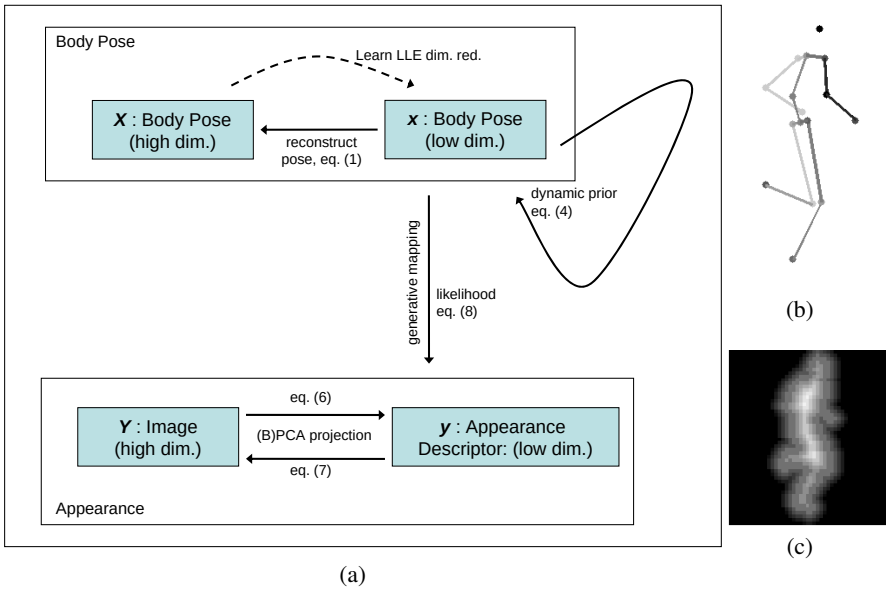


Fig. 1. a) An overview of the tracking framework. Solid arrows represent signal flow during inference, the dashed arrow stands for the nonlinear dimensionality reduction during training. The figure refers to equations in Section 2. b) Body pose representation as a number of 3d joint locations. c) Distance 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.

at different speeds. Walking and running training examples were separately processed to train activity specific models.

2.1 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, 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) [5], 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 each data point in a space of desired low dimensionality. 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 (RVM). 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. 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 $f_{switch}^{r\rightarrow w}(x^r)$ and $f_{switch}^{w\rightarrow r}(x^w)$. This can easily be generalised to more action categories and leads to the following motion model, where the state space from eq. (4) is augmented by a discrete state variable a_t.

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

Here, the motion model for the case of activity switching $p^{a_{t-1}\rightarrow a_t}(x_t|x_{t-1})$ is modelled as a normal distribution around the pose predicted by the regressor $f_{switch}^{a_{t-1}\rightarrow a_t}$. The probabilities that an activity transition does or does not occur are denoted p_{switch} and $p_{noswitch}$.

2.2 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* $dt(Y)$ [16] of segmented figures with a subsequent linear PCA dimensionality reduction step (see Figure 1c), and a representation obtained by applying *Binary PCA (BPCA)* [17] 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 [17]). 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 probability maps. From the descriptors we compute probability maps via the sigmoid function $\sigma(.)$.

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

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.

3 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 distinguish subjects from each other and track them through occlusions. 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. Our experiments have shown that a strong dynamical model is necessary to avoid confusion between these posterior modes and reduce ambiguities. The remaining posterior multimodalities 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* [18], 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 eq. (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. 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 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.

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 eq. (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 eq. (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 [19] 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 matrix Σ_a. 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.

3.1 Activity Switching

When turning to the multi activity tracking case, the sample proposal function is adapted according to eq. (5). A sample i undergoes an activity switch with probability p_{switch}. We currently consider two activities, walking and running, therefore we set $p_{switch}^{w \to r} = p_{switch}^{r \to w} = 1 - p_{noswitch}$. The scheme can be extended to more activity categories. 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 information, 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_a^w resp. f_a^r in eq. (8)).

3.2 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 $p(X_t | Y_{1:t})$. Often we are interested in a consistent trajectory through the entire image sequence, i.e. in the maximum of the posterior $p(X_{1:T} | Y_{1:T})$ over the poses of *all* time steps, given *all* observations. In other words, we are interested in the value for $X_{1:T}$ with maximal probability rather than marginals for each X_t.

In our framework this is achieved by a postprocessing algorithm that finds optimal paths through the set of samples. We use the Max-Product algorithm resp. its numerically more stable counterpart, the Min-Sum algorithm that operates with negative logarithms instead of probabilities (see [20] or [21] for belief propagation algorithms). These algorithms are discrete by nature, i.e. each node of the Markov chain (each time step) has a number of discrete states that in our case is equal to the number of samples N of the 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.

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 *evidence* 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. Since the global function to optimise is the same as during the recursive tracking, the same equations can be reused. The *evidence* functions

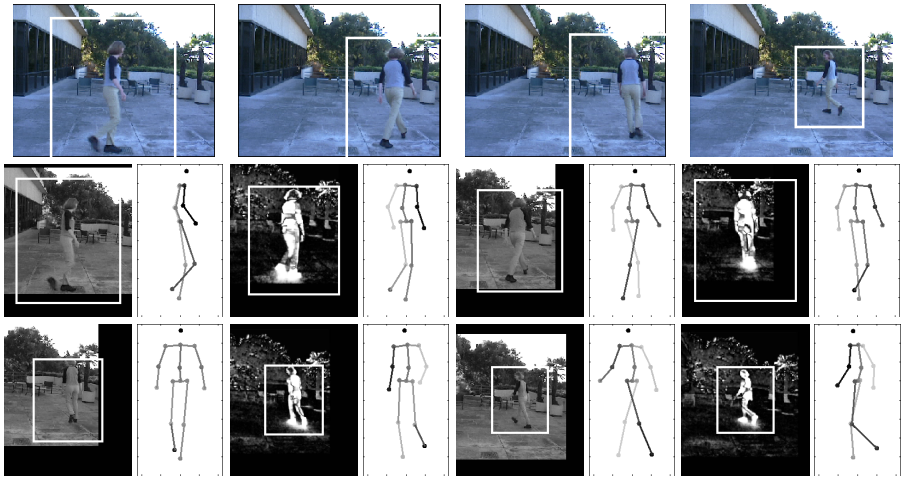


Fig. 2. Circular walking sequence from [22]. The figure shows full frames (top), and cutouts with bounding box in original or segmented input images and estimated poses. Darker limbs are closer in depth.

$v(\Theta_t)$ are computed according to eq. (11). In fact we can directly reuse the sample weights computed during tracking. The *compatibility* between neighbouring nodes is given by eq. (10). The Max-Product resp. Min-Sum algorithm performs inference in this chain graph by propagating local messages (*beliefs*) between neighbouring nodes.

4 Experiments

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 framerate 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 2000 body poses of each activity. All the kernel regressors were trained using the Relevance Vector Machine algorithm [23], 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.

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 Section 3.2, based on the samples from the recursive Bayesian sampling algorithm.

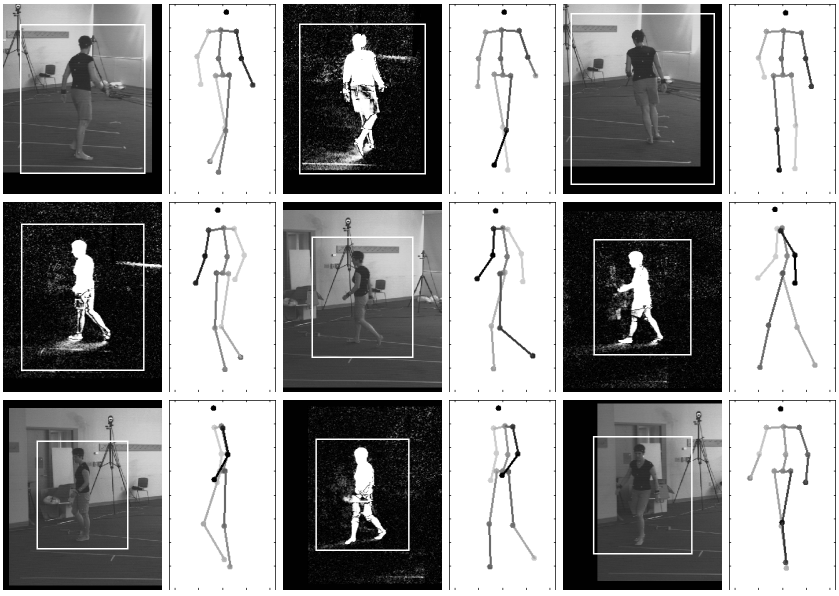


Fig. 3. Circular walking sequence from [24], original resp. segmented input images with estimated bounding boxes, and estimated poses.

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. 2) shows tracking on a standard test sequence[1] from [22], where a person walks in a circle. We segmented the images using 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 3 shows another publicly available sequence[2]. Here we used only one camera, this sequence has been mainly used for multi-camera tracking (e.g. [24, 9]).

Figure 4 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 on the bottom right shows the estimated

[1] http://www.nada.kth.se/ hedvig/data.html
[2] http://www.cs.brown.edu/ ls/

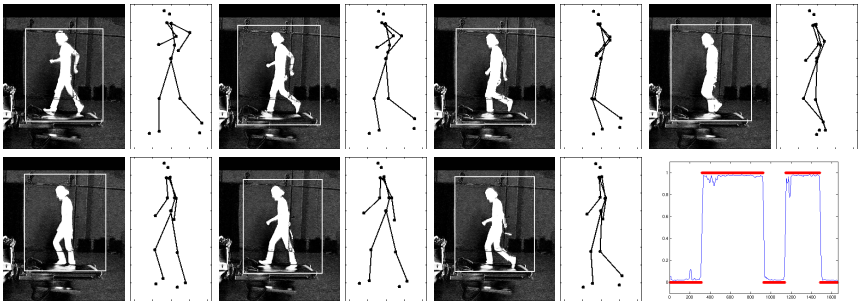


Fig. 4. Transition from running to walking. The original sequence is 1660 frames long, here we show selected frames from the transition phase between frame 921 and frame 936. The figure on the bottom right shows the estimated activities; the blue curve shows the continuous probability that we observe running rather than walking over the entire sequence, the red bars indicate the activity that has been inferred by the global optimisation.

running probabilities throughout the sequence. Even for humans, it is not obvious to identify the exact moment of activity change, there is typically a transition phase of approx. 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 Figures 5 and 6 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. In Figure 5 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 Figure 6 suddenly starts to run when crossing a street. The activity switch is reliably detected, as can be seen in the activity plot on the bottom right.

5 Summary

We presented an 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 different challenging video sequences of low resolution with noisy segmentation.

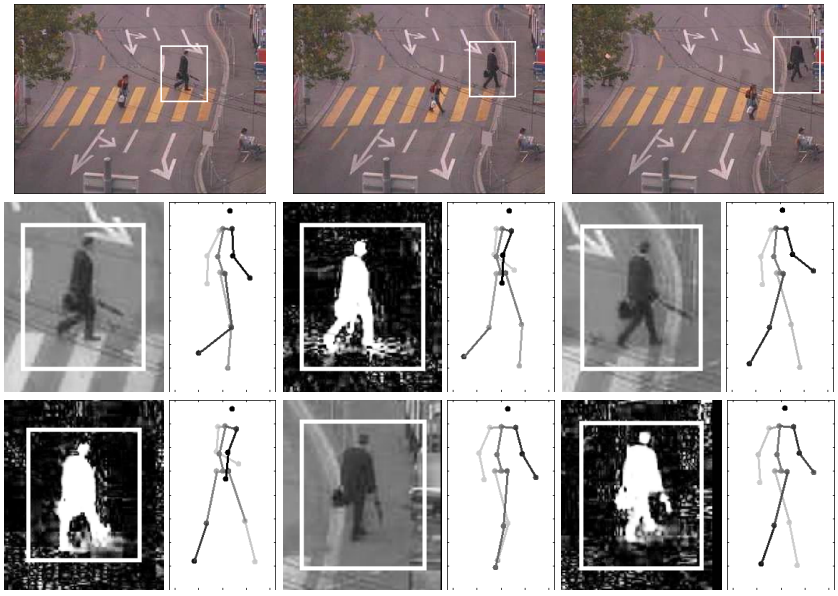


Fig. 5. Real traffic scene with low resolution input images, noisy segmentation, disturbing objects (umbrella, bag), and varying viewangle. Original frames (top) and cutouts.

Acknowledgements

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

References

1. Rosales, R., Sclaroff, S.: Learning body pose via specialized maps. NIPS (2001)
2. Thayananthan, A., Navaratnam, R., Stenger, B., Torr, P., Cipolla, R.: Multivariate relevance vector machines for tracking. Ninth European Conference on Computer Vision (2006)
3. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Discriminative density propagation for 3d human motion estimation. CVPR (2005)
4. Agarwal, A., Triggs, B.: Monocular human motion capture with a mixture of regressors. IEEE Workshop on Vision for Human-Computer Interaction at CVPR (2005)
5. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science, v.290 no.5500 , Dec.22, 2000, pp.2323–2326 (2000)
6. Forsyth, D.A., Arikan, O., Ikemoto, L., J. O' Brien, D.R.: Computational studies of human motion: Part 1. Computer Graphics and Vision Volume 1 Issue 2/3 (2006)
7. Agarwal, A., Triggs, B.: 3d human pose from silhouettes by relevance vector regression. In CVPR (2004)
8. Grauman, K., Shakhnarovich, G., Darrel, T.: Inferring 3d structure with a statistical image-based shape model. In ICCV (2003)
9. Sun, Y., Bray, M., Thayananthan, A., Yuanand, B., Torr, P.: Regression-based human motion capture from voxel data. Proceedings British Machine Vision Conference (2006)
10. Lim, H., Camps, O.I., Sznaier, M., Morariu, V.I.: Dynamic appearance modeling for human tracking. In: Conference on Computer Vision and Pattern Recognition. (2006) 751–757

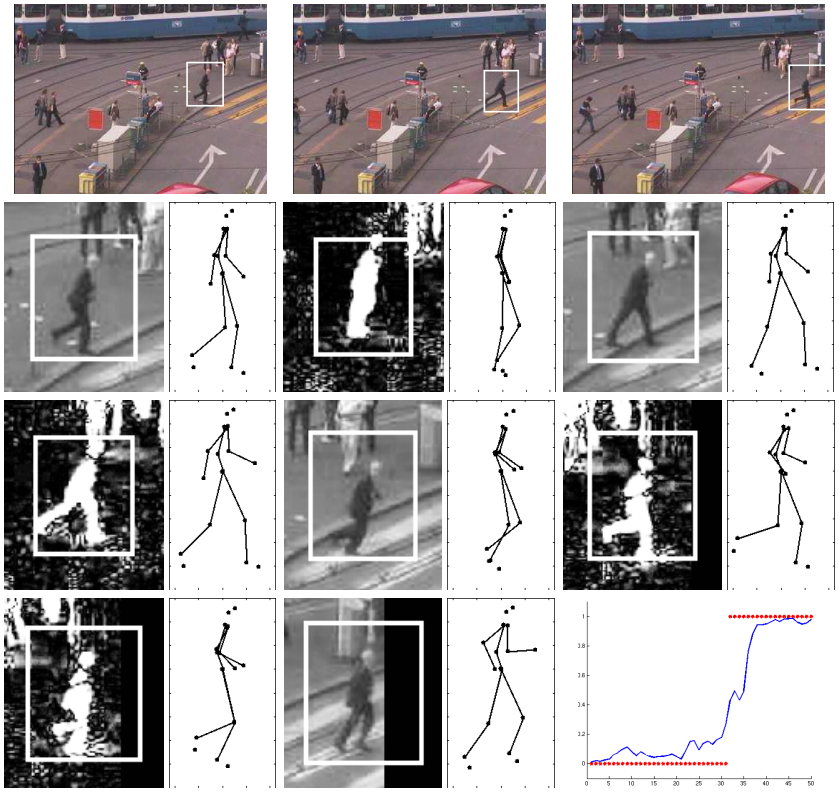


Fig. 6. Real traffic scene with a transition from walking to running. Full Frames (top) and cutouts with estimated poses. The figure on the bottom right shows the probability of action category *running* (blue, solid line), and the activity inferred by the global optimisation (red dots).

11. Elgammal, A., Lee, C.S.: Inferring 3d body pose from silhouettes using activity manifold learning. CVPR (2004)
12. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models. In: Advances in Neural Information Processing Systems 18. (2006) 1441–1448
13. Sminchisescu, C., Jepson, A.: Generative modeling for continuous non-linearly embedded visual inference. International Conference on Machine Learning, ICML (2004)
14. Li, R., Yang, M.H., Sclaroff, S., Tian, T.P.: Monocular tracking of 3d human motion with a coordinated mixture of factor analyzers. ECCV (2) (2006) 137–150
15. Isard, M., Blake, A.: A mixed-state CONDENSATION tracker with automatic model-switching. In: ICCV. (1998) 107–112
16. Bailey, D.G.: An efficient euclidean distance transform. IWCIA (2004)
17. Zivkovic, Z., Verbeek, J.: Transformation invariant component analysis for binary images. CVPR (1) 2006: 254-259 (2006)
18. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. Int. J. Computer Vision (1998)
19. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math Soc. (1943)
20. Kschischang, F., Frey, B.J., Loeliger, H.A.: Factor graphs and the sum-product algorithm. IEEE Trans. Info. Theory 47, 498–519 (2001)

21. Yedidia, J., Freeman, W., Weiss, Y.: Understanding belief propagation and its generalizations. Technical Report TR-2001-22, MERL (2002)
22. Sidenbladh, H., Black, M., Fleet, D.: Stochastic tracking of 3d human figures using 2d image motion. In ECCV (2000) 702–718
23. Tipping, M.: The relevance vector machine. In: NIPS. (2000)
24. Sigal, L., Bhatia, S., Roth, S., Black, M., Isard, M.: Tracking loose-limbed people. CVPR (2004)