

Learning Generative Models for Monocular Body Pose Estimation

Tobias Jaeggli¹, Esther Koller-Meier¹, 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 optimization.

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 single-valued 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 parametrisations. In order 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 single-valued and thus modelled with a single linear regressor. The one-to-many 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 losing 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. This approach does 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 parametrisation, with learned mappings back to the original pose space and into the appearance space. In this section all elements of the framework will be described in detail. Our models were trained on real motion capture data sets of

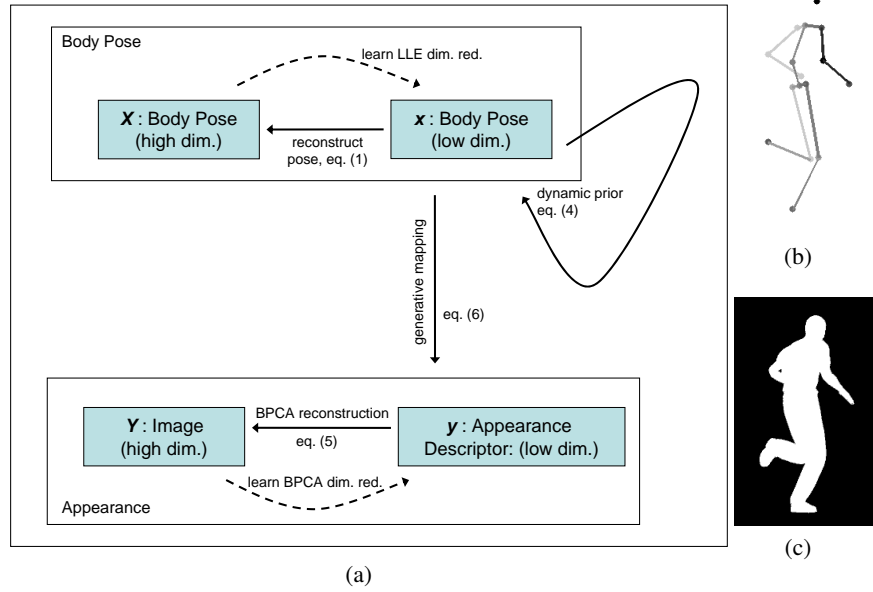


Fig. 1. a) An overview of the tracking framework. Solid arrows represent signal flow during inference, the dashed arrow stands for LLE resp. BPCA dimensionality reduction during training. The figure refers to equations in Section 2. b) Body pose representation as a number of 3d joint locations. c) Corresponding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

$$X = f_p(x) = W_p \Phi_p(x) \quad (1)$$

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,

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

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).

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

with C the number of mixture components. We obtain the following formulation for the temporal prior by combination with the *dynamic* prior $p_d(x_t|x_{t-1})$.

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

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 $\sigma(Y_c)$.

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

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 ,

$$f_a(x, \omega) = W_a \Phi_a(x, \omega) \quad (6)$$

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.

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

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.

$$Obs(\mathbf{p}) = p(\mathbf{p} = fg|Y_{obs})$$

$$BC(x, \omega, Y_{obs}) = \sum_{\mathbf{p}} \sqrt{Seg(\mathbf{p})Obs(\mathbf{p})} \quad (8)$$

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 the observation likelihood

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

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 ,

$$\Theta_t = [\omega_t, u_t, v_t, w_t, h_t, x_t], \quad (10)$$

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 $\theta = [\omega_t, u_t, v_t, w_t, h_t]$.

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

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.

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

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 segmentation, very low resolution, and varying viewpoints.

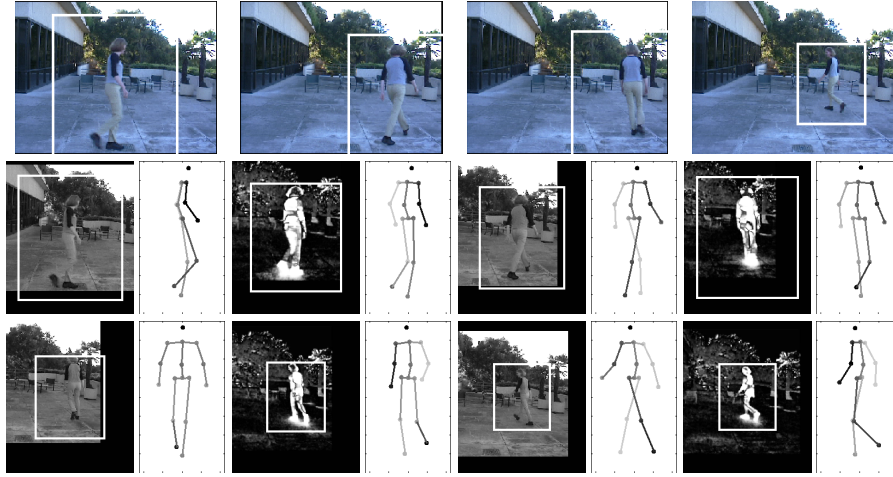


Fig. 2. Circular walking sequence from [5]. The figure shows full frames (top), and cutouts with bounding box in original or segmented input images and estimated poses. Darker limbs are closer in depth.

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.

The first experiment (Fig. 2) shows tracking on a standard test sequence³ from [5], where a person walks in a circle. We segmented the images using background sub-

³ <http://www.nada.kth.se/~hedvig/data.html>

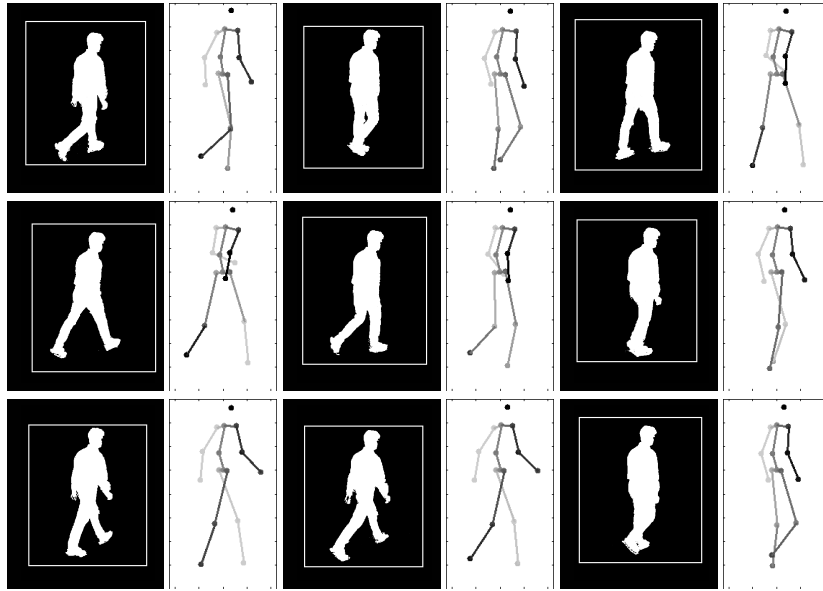


Fig. 3. Diagonal walking sequence. Estimated bounding boxes and poses. The intensity of the stick figure limbs encodes depth; lighter limbs are further away.

traction, 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 with unreliable segmentations and multimodal per-frame likelihoods.

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

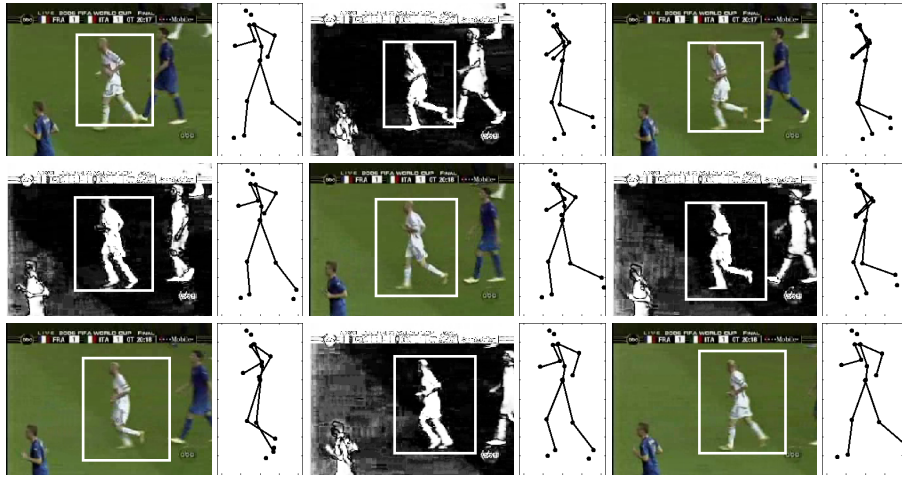


Fig. 4. An extract from a soccer game. The figure shows original and segmented images and with estimated bounding boxes, and estimated 3d poses.

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)
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

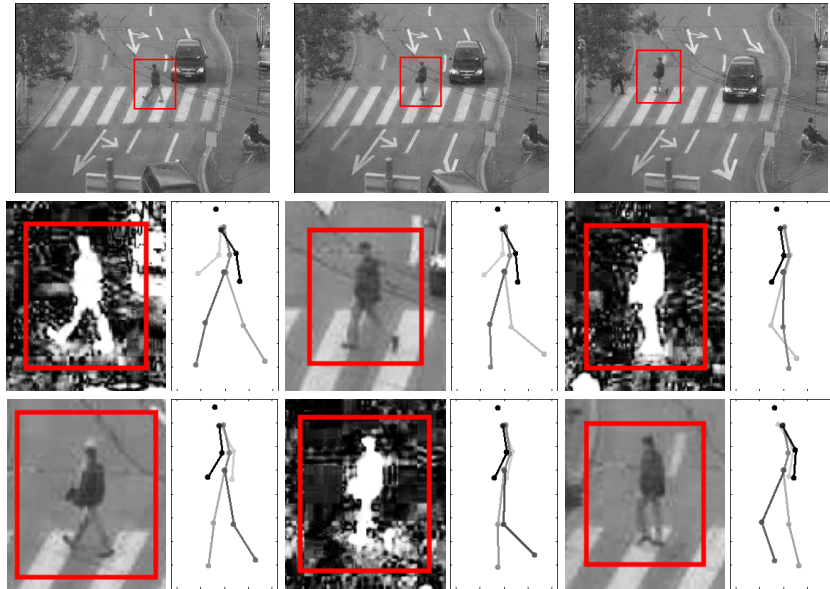


Fig. 5. Traffic scene with low resolution images and noisy segmentation.

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., Sznaiier, 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)