Bridging the Gap between Detection and Tracking for 3D Monocular Video-Based Motion Capture *

Andrea Fossati, Miodrag Dimitrijevic, Vincent Lepetit, Pascal Fua
Computer Vision Laboratory http://cvlab.epfl.ch
École Polytechnique Fédérale de Lausanne (EPFL) *1015 Lausanne, Switzerland*
{andrea.fossati, miodrag.dimitrijevic, vincent.lepetit, pascal.fua}@epfl.ch

Abstract

We combine detection and tracking techniques to achieve robust 3–D motion recovery of people seen from arbitrary viewpoints by a single and potentially moving camera. We rely on detecting key postures, which can be done reliably, using a motion model to infer 3–D poses between consecutive detections, and finally refining them over the whole sequence using a generative model.

We demonstrate our approach in the case of people walking against cluttered backgrounds and filmed using a moving camera, which precludes the use of simple background subtraction techniques. In this case, the easy-to-detect posture is the one that occurs at the end of each step when people have their legs furthest apart.

1. Introduction

Recent approaches to modeling people's 3–D motion from video sequences can be roughly classified into those that detect specific postures in individual frames and those that track the motion from frame to frame given an initial pose. The first category usually involves matching against a large database and is becoming increasingly popular, but requires very large training data sets to be effective. The second category involves predicting the pose in a frame given the pose computed in previous frames, which can easily fail if errors start accumulating in the prediction, causing the estimation process to diverge.

Neither technique is clearly superior to the other, and both are actively investigated. In this paper, we show that they can be combined to accurately reconstruct the 3–D motion of people seen from arbitrary viewpoints using a single, and potentially moving, camera. At the heart of our approach is the fact that human motions often contain characteristic postures that are relatively easy to detect. Given two consecutive such postures, modeling intermediate poses becomes an interpolation problem, which is much easier to perform reliably than open-ended tracking.

More specifically, we demonstrate our approach in the case of people walking along arbitrary trajectories. In this case, the easy-to-detect posture is the one that occurs at the end of each step when people have their legs furthest apart. We therefore use a chamfer-based method [7] that was designed to detect this posture from any viewpoint, even when the background is cluttered and background subtraction is impractical because the camera moves as is the case in the first row of Fig. 1. Because the detected postures are projections of 3–D models, we can map them back to full 3–D poses and use them to select and warp motions from a training database that closely match them. This yields initial pose estimates such as those of the second row of Fig. 1. It lets us create the synthetic images we would see if the person truly were in those positions. These images are depicted by the figure's third row and we refine the pose until they match the real ones. This yields the results depicted by the two last rows of Fig. 1.

The importance of combining detection and tracking to achieve robustness has long been known [5, 12] and manually introducing a few 3D keyframes in a tracking algorithm has been shown to be effective [6]. More recently, a fully automated approach to combining tracking and detection has been shown to be very effective at following multiple people over very long sequences in [15] in 2–D. This is achieved by detecting people in canonical poses and tracking them from there, which still has the potential to diverge. By contrast, interpolating between detected silhouettes prevents this and yields 3–D reconstructions.

We chose walking to demonstrate our approach because we had access to both the appropriate motion database and silhouette detection technique. The approach, however, is general because most human motions include very charac-

*This work has been partially funded by the VISIONTRAIN RTN-CT-2004-005439 Marie Curie Action within the EC's Sixth Framework Programme. The text reflects only the authors' views and the Community is not liable for any use that may be made of the information contained therein.

Figure 1. Our approach. **First row:** Input sequence acquired using a moving camera with silhouettes detected at the beginning and the end of the walking cycle. The projection of the ground plane is overlaid as a blue grid. **Second row:** Projections of the 3–D poses inferred from the two detections. **Third row:** Synthesized images that are most similar to the input. **Fourth row:** Projections of the refined 3–D poses. **Fifth row:** 3–D poses seen from a different viewpoint.

teristic postures that are easier to detect than completely arbitrary ones. Athletic motions are a good example of this. Canonical postures can be detected when a tennis player hits the ball with a forehand, a backhand, or a serve [20]. The same can be said when a golfer begins the upswing, transitions from upswing to downswing, and completes the motion. In a work environment, there also are very characteristic poses between which people alternate, such as sitting at their desk and walking through the door. In short, canonical postures are common. This is important because one of the limitations of state-of-the-art detection-based approaches to 3–D motion reconstruction is that huge training databases would be required to detect all possible postures. By contrast, if one only needs to detect a few easily recognizable postures, much smaller databases should suffice.

2. Related Work

Existing approaches to video-based 3D motion capture remain fairly brittle for many reasons: Humans have a complex articulated geometry overlaid with deformable tissues, skin and loose clothing. They move constantly, and their motion is often rapid, complex, and self-occluding. Furthermore, the 3D body pose is only partially recoverable from its projection in one single image. Reliable and robust 3D motion analysis therefore requires good tracking across frames, which is difficult because of the poor quality of image-data and frequent occlusions. Recent approaches to handling these problems can roughly be classified into those that

- *Detect:* This implies recognizing postures from a single image by matching it against a database and has become increasingly popular recently [22, 1, 9, 13, 10,

7, 8] but requires very large sets of examples to be effective. Moreover this often relies on background subtraction, which requires static cameras.

- *Track:* This involves predicting the pose in a frame given observation of the previous one. This requires an initial pose and can easily fail if errors start accumulating in the prediction, causing the estimation process to diverge. This is usually mitigated by introducing sophisticated statistical techniques for a more effective search [5, 3, 4, 25, 26, 19] or by using strong dynamic motion models as priors [16, 14, 17, 2, 24, 21].

Neither technique has been proved to be superior, and both are actively studied and sometimes combined: Manually introducing a few 3D keyframes is known to be a powerful way to constrain 3D tracking algorithms [6, 11]. In the 2D case, it has recently been shown that this can be done in a fully automated way to track multiple people in extremely long sequences [15]. This involves tracking forwards and backwards from individual and automatically detected canonical poses. While effective, this approach to tracking still has the potential to diverge. In this paper, we avoid this problem and go to full 3D by observing that automated canonical pose detections can be linked into complete trajectories, which let us first recover rough 3D poses by interpolating between these detections and then refining them by using a generative model over full sequences. A similar approach has been proposed for 3D hand tracking [23] but makes much stronger assumptions than ours, requiring a perfect hand view with an easy to remove background.

3. Approach

Here, we first give a short overview of the complete approach, before going into more details in the following subsections.

3.1. Overview

To initialize our system, we use a template-based approach [7] to detect people at the moment of the walking cycle when their legs are furthest apart, as shown in the first row of Fig. 1. The templates consist of consecutive 2D silhouettes obtained from 3D motion capture data seen from six different camera views and at different scales. This way the motion information is incorporated into the templates and helps distinguish actual people who move in a predictable way from static objects whose outlines roughly resemble those of humans. For each detection, the system returns a corresponding 3D pose estimate.

In theory, a person should be detected at the beginning of each walking cycle, which the template-based algorithm does with a low error rate. The few false positives tend to correspond to actual people but detected at somewhat inaccurate scales or orientations and false negatives occur when the person faces the camera head-on and the characteristic pose we are looking for becomes hard to distinguish from the others. We have therefore designed an approach based on dynamic-programming that links detections into consistent trajectories, even though a few may have been missed. Since the camera may move, we perform this computation in the ground plane, which we relate to the image plane via a homography that is recomputed from frame to frame.

Finally, we use consecutive detections on individual trajectories to select and time-warp motions from a training database obtained via optical motion capture. As shown in the second row of Fig. 1, this gives us a rough estimate of the body's position and configuration in each frame between detections. To refine this initial estimate, and since the camera may move from frame to frame, we first compute homographies between consecutive frames and use them to synthesize a background scene from which the moving person has been almost completely removed. We then learn an appearance model from the detections and use it in conjunction with the synthetic background to produce new images, which lets us refine the body position by minimizing the difference between the original and synthetic images. This yields the refined poses depicted by the bottom three rows of Fig. 1.

We describe below our approach to linking sparse detections into complete trajectories and, then, to inferring 3D poses for the whole sequence. For additional details on detection itself we refer the interested reader to [7].

3.2. From Detections to Trajectories

In our scheme, people should be detected at every walking cycle but are occasionally missed. To link these sparse detections into a complete trajectory, we have implemented a Viterbi-style algorithm. Note that these detections include not only an image location but also the direction the person faces, which is an important clue for linking purposes.

Ground Plane Registration. Since the camera may move, we work in the ground plane, which we relate to each frame by a homography that is computed using standard techniques [18]. In practice, we manually indicate the ground plane in one frame and compute an initial homography between it and the world ground plane. Then, we detect interest points in both the reference frame and the next one, match them, use the resulting correspondences to compute the next homography, and repeat this process for all subsequent frames.

Formalizing the Problem. The homographies let us compute ground plane locations and one of eight possible

orientations for all detections, which we then need to link while ignoring potential misdetections. To this end, we define a hidden state at time t as the oriented position of a person on the ground plane $L_t = (X, Y, O)$, where t is a frame index, (X, Y) are discretized ground plane coordinates, and O is one of eight possible orientations.

We introduce the maximum likelihood estimate of a person's trajectory ending up at state i at time t

$$\Psi_t(i) = \max_{l_1,...,l_n} P(I_1, L_1 = l_1, ..., I_n, L_n = l_n) \ , \quad (1)$$

where I_j represents the j^{th} frame of the video sequence. Casting the computation of Ψ in a dynamic programming framework requires introducing probabilities of observing a particular image given a state and of transitioning from one state to the next.

We therefore take b_{it}, the probability of observing frame I_t given hidden state i, to be

$$b_{it} = P(I_t|L_t = i) \sim \frac{1}{d_{\text{chamfer}}} \ , \quad (2)$$

where d_{chamfer} is a weighted average of the chamfer distances between projected template contours and actual image edges. This makes sense because the coefficients used to weight the contributions are designed to account for the relevance of different silhouette portions in a Bayesian framework [7].

We also introduce the probability of transition from state j at time t' to state i at time t

$$a_{ji}^{\Delta t} = P(L_t = i|L_{t'} = j), \Delta t = t - t'. \quad (3)$$

Since we only detect people when their legs are spread furthest apart, we can only expect a detection approximately every $N_c = 30$ frames for an average v = 5 km/h walking speed in a 25 Hz video. This implies an average distance $d_c = \frac{vN_c}{25}$ between detections. We therefore assume that $a_{ji}^{\Delta t}$ for state $i = (X, Y, O)$ follows a Gaussian distribution centered at (X_μ, Y_μ) such that

$$\sqrt{(X - X_\mu)^2 + (Y - Y_\mu)^2} = d_c \ , \quad (4)$$

and positioned in the direction 180° opposite to the orientation O, as depicted by point A in Fig. 2. This Gaussian covers only the hidden states with orientation equal to O. The other previous states from which a transition may occur are those with orientations $O+\pi/4$ and $O-\pi/4$, which are covered by two neighboring Gaussians, as depicted by points B and C in Fig. 2.

Linking Sparse Detections. Given the probabilities of Eq. 2 and 3, if we could expect a detection in every frame, linking them into complete trajectories could be done using the Viterbi algorithm to recursively maximize the Ψ_t maximum likelihood of Eq. 1.

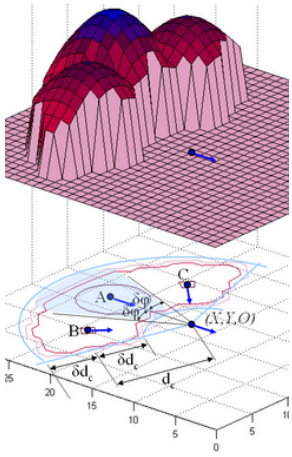

Figure 2. Transitional probabilities for hidden state (X, Y, O). They are represented by three Gaussian distributions corresponding to three possible previous orientations. Each Gaussian covers a 2D area bounded by two circles of radii $d_c - \delta d_c$ and $d_c + \delta d_c$, where δd_c represents an allowable deviation from the mean, and by two lines defined by tolerance angle $\delta\varphi$.

However, since we can only expect a detection approximately every $N_c = 30$ frames, we allow the model to change state directly from $L_{t'}$ at time t' to L_t at time t $(t' < t)$, $N_c - \delta t < t - t' < N_c + \delta t$ and skip all frames in between. δt is a frame distance tolerance that we set to 10 in our implementation.

This lets us reformulate the maximization problem of Eq. 1 as one of maximizing

$$\begin{aligned} \Psi_t(i) &= \max_{l_{t_1},...,l_{t_n}} P(I_{t_1}, L_{t_1} = l_{t_1}, ..., I_{t_n}, L_{t_n} = l_{t_n}) \ , \\ &= b_{it} \max_{j,\Delta t}(a_{ji}^{\Delta t} \Psi_{t-\Delta t}(j)) \ , \end{aligned} \quad (5)$$

where $t_1 < t_2 < ... < t_n$, n are the indices of frames I in which at least one detection occurred and $N_c - \delta t < \Delta t < N_c + \delta t$.

This formulation lets us initially retain for each detection several hypotheses with different orientations and allow the dynamic programming algorithm to select those that provide the most likely trajectories according to the probabilities of Eq. 2 and 3. If a detection is missing, the algorithm simply bridges the gap using the transition probabilities only. For the sequences of Fig. 5, this yields the results depicted by Fig. 6.

3.3. Predicting 3D Poses between Detections

A complete trajectory computed as discussed above includes rough estimates of the body's position, orientation, and 3D pose parameterized by a set of joint angles for the

frames in which the key posture was detected. We now turn to inferring approximate body poses for all the frames between detections and will discuss in the following subsection how we refine them by going back to the actual images.

Let us first consider the case where the key postures at the start and end of the walking cycle are both detected, which in practice is the most frequent one. We represent the human body as a set of cylinders attached to an articulated 3–D skeleton and its *pose* is given by the position and orientation of its root node, defined at the sacroiliac, and a set of joint angles. We use straightforward spline interpolation to predict positions and orientations between the two detections. To predict the joint angle configurations, we take advantage of the fact that the templates used to detect people were created using a motion capture database to which we have access. Given a detected silhouette, we can therefore select in the database the corresponding motion $\mathcal{M}$. Since the number of poses in $\mathcal{M}$ is not necessarily equal to the number of frames between the two detections, we resample $\mathcal{M}$ in time. To this end, each pose in $\mathcal{M}$ is represented by its coordinates into an eigen-space of dimension 3 computed from the database, and the resampling is performed in this low-dimensional space for better stability. The joint angles are then finally retrieved by simply back-projecting the resampled poses.

This procedure is very simple and naturally extends to the case where a key posture has been missed, which can be easily detected by comparing the number of frames between consecutive detections and the median value for the whole sequence. In this case, a longer motion must be created by repeating $\mathcal{M}$ several times —usually 2, and never more than 3 in our experiments— depending on the number of frames between detections. This new motion is then resampled as before. Obviously the predictions then lose in accuracy, but they usually remain precise enough to retrieve the correct poses thanks to the refinement process described in the following section.

3.4. Refining the Predicted Poses

We now turn to refining the initial set of 3–D poses using a generative approach: For a given body pose, we generate the synthetic image we would see if the person truly were in that position and compare to the original one. Minimizing the difference between real and synthetic image then lets us refine the poses in each individual frame.

This is standard but we add an important novel element which we have found to be key to obtaining good results, as illustrated by Fig. 3: We not only create an appearance model for the person but also for the background so that the synthetic images we produce include both. This effectively constrains the projections of the reconstructed model to project at the right place and allows us to recover the correct pose even when the initial guess is far from it.

Figure 3. Refinement process. The images are from left to right the input image, the initialization given by the interpolation process, the refinement obtained without using the background generation, and finally the refinement obtained as proposed. Whole parts of the body can be missed when the background is not exploited.

Fig. 4 depicts our approach to computing the background images. Given an image of the sequence, we treat it as a reference and consider the few images immediately before and after. We compute homographies between the reference and all other images [18], which is a reasonable approximation of the frame-to-frame deformation because the time elapsed between successive frames is short and lets us warp all the images into the reference frame. Subsequently, by taking the median of the values for each pixel in HSV color space, we obtain background images with few artifacts.

Given these generated background images, we project our human body model according to the pose we want to evaluate. As discussed in Section 3.3, individual limbs are modeled as cylinders to which we associate a color by averaging pixel intensities in the projected area of the limb in the frames where the silhouette was detected. We project the body model onto the generated background image to obtain a synthetic image, such as those depicted by the third row of Fig. 1. A pose can then be evaluated by computing the sum-of-squared-differences (SSD) between this synthetic image and the actual one. Incorporating both foreground and background into our synthetic images makes the measure reliable enough so that we did not have to use a robust estimator for our experiments. However, because it produces many local minima when the pose changes, we use a simple stochastic optimization technique that samples the pose space around the predicted pose in the low dimensional space and retain the sample that yields the smallest SSD. Because we only search for poses around the predicted ones, we still benefit from the constraints provided by interpolating between key-poses.

4. Results

The algorithm presented in the previous sections allows us to robustly and automatically retrieve the 3–D pose of a walking person, different from the subjects we used to create the database, without drift. We demonstrate this using sequences acquired with a moving camera and where the person may be seen from very different angles.

Figs. 7, 8, and 9 depict excerpts of such sequences, which we provide as supplemental material.

Figure 4. Synthesizing a background image. **First row:** The rightmost image is the reference image whose background we want to synthesize. The other 4 are those before and after it in the sequence. **Second row:** The same four images warped to match the reference image. Computing the median image of these and the reference image yields the rightmost image, which is the desired background image.

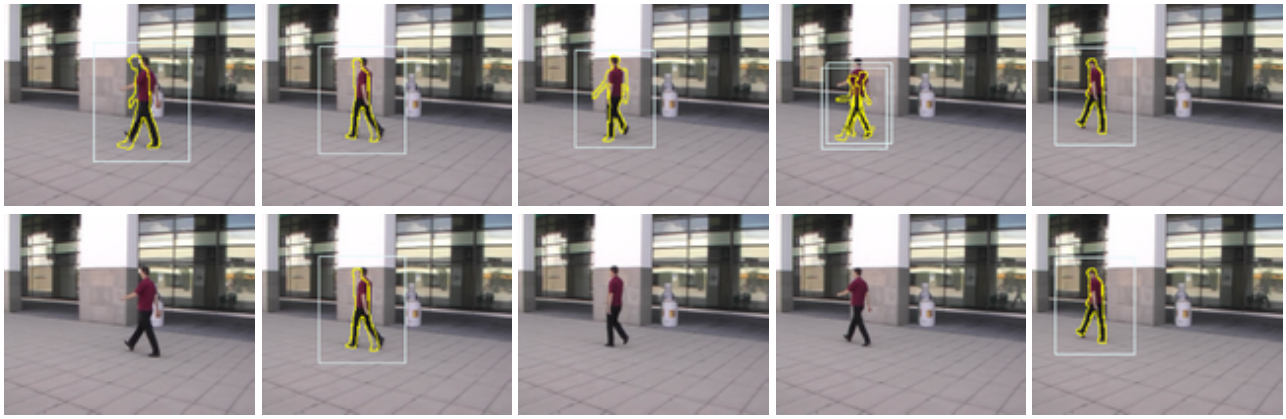

Figure 5. Filtering silhouettes with temporal consistency on an outdoor sequence acquired by a moving camera. **First row:** Detection hypotheses. **Second row:** Detections after filtering out the detection hypotheses that do not lie on the recovered most probable trajectory.

In the sequence of Fig. 7 the camera translates and the subject is seen first from the side and progressively from the back as he becomes smaller and smaller. In Fig. 9, we highlight the robustness of our approach to missed detections: Using only one detection out of every two does not substantially degrade the performance.

5. Conclusion

The walking motion contains a characteristic posture that is relatively easy to detect. We have exploited this fact to formulate 3–D motion recovery from a single video sequence as an interpolation problem. This is much easier to achieve than open-ended tracking and we have shown that it can be solved using straightforward minimization.

This approach is generic because most human motions also feature canonical poses that can be easily detected. This is significant because it means that we can focus our future efforts on developing methods to reliably detect these canonical poses instead of all poses, which is much harder. In future research, we will therefore extend our approach to other motions, such as running, swinging a golf club, or playing tennis.

References

[1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *Conference on Computer Vision and Pattern Recognition*, 2004.

[2] A. Agarwal and B. Triggs. Tracking articulated motion with piecewise learned dynamical models. In *European Conference on Computer Vision*, Prague, May 2004.

[3] K. Choo and D. Fleet. People tracking using hybrid monte carlo filtering. In *International Conference on Computer Vision*, Vancouver, Canada, July 2001.

[4] A. J. Davison, J. Deutscher, and I. D. Reid. Markerless motion capture of complex full-body movement for character animation. In *Eurographics Workshop on Computer Animation and Simulation*. Springer-Verlag LNCS, 2001.

[5] J. Deutscher, A. Blake, and I. Reid. Articulated Body Motion Capture by Annealed Particle Filtering. In *Conference on Computer Vision and Pattern Recognition*, pages 2126–2133, Hilton Head Island, SC, 2000.

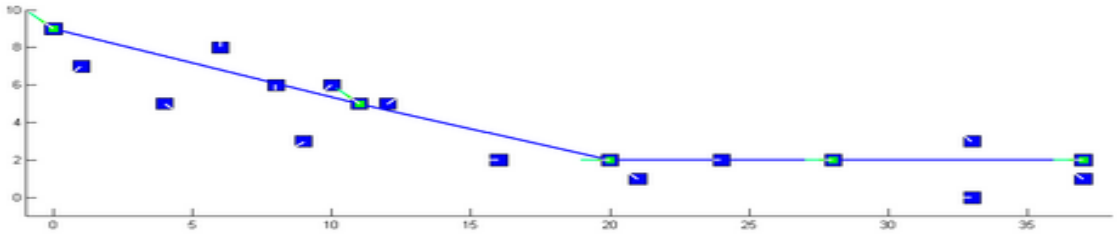

Figure 6. Recovered trajectory for the sequence depicted by Fig. 5. Dark blue squares represent detection hypotheses and bright short lines inside them represent the detection orientations. Smaller light green squares and lines represent the retained detections and their orientations respectively. These detections form the most probable trajectory depicted by dark blue lines.

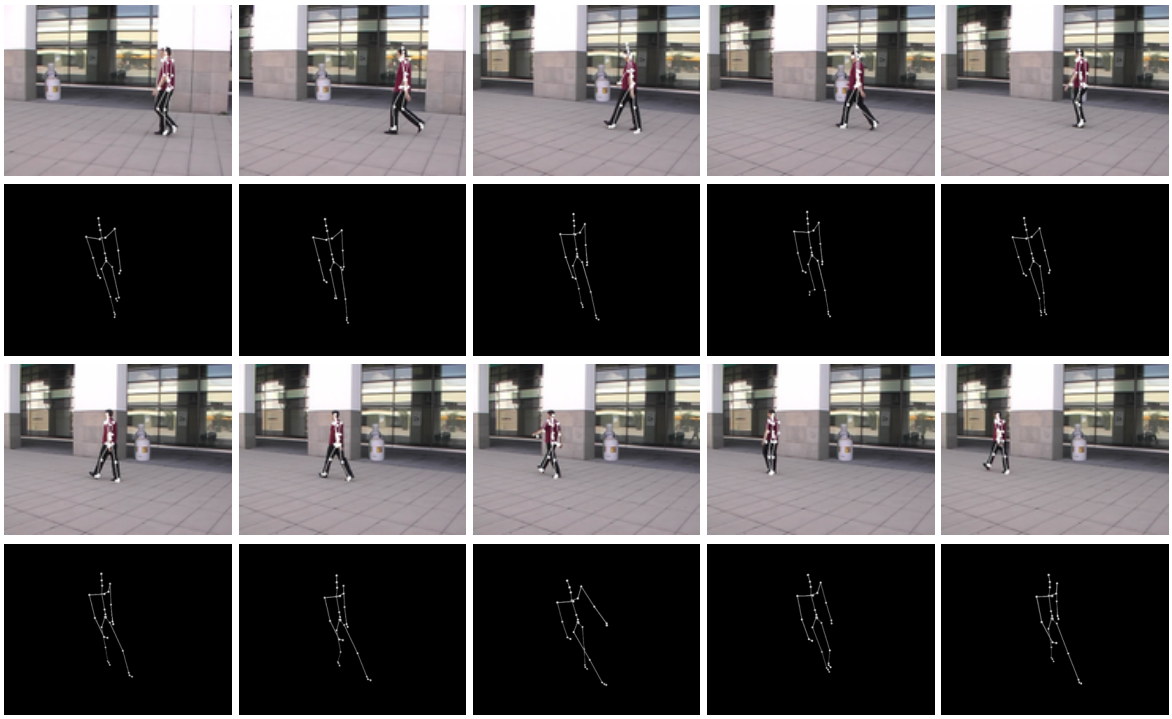

Figure 7. Results under viewpoint changes on a sequence shot by a moving camera. Both 2D projection and 3D model seen from a different viewpoint are provided. The 3–D pose is correctly estimated over the sequence, even when the person goes far away and turns with respect to the camera. Note that the extremely similar poses in which is very hard to distinguish which leg is in front of which leg are successfully disambiguated by our algorithm. The sequence, lasting around 150 frames, is given as supplemental material.

[6] D. DiFranco, T. Cham, and J. Rehg. Reconstruction of 3-D Figure Motion from 2-D Correspondences. In *Conference on Computer Vision and Pattern Recognition*, 2001.

[7] M. Dimitrijevic, V. Lepetit, and P. Fua. Human Body Pose Detection Using Bayesian Spatio-Temporal Templates. *Computer Vision and Image Understanding*, 104(2-3):127–139, 2006.

[8] E.-J-Ong, A. S. Micilotta, R. Bowden, and A. Hilton. Viewpoint invariant exemplar-based 3–d human tracking. *Computer Vision and Image Understanding*, 104(2–3):178–189, 2006.

[9] A. Elgammal and C. Lee. Inferring 3D Body Pose from Silhouettes using Activity Manifold Learning. In *Conference on Computer Vision and Pattern Recognition*, 2004.

[10] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Conference on Computer Vision and Pattern Recognition*, volume 1, San Diego, CA, June 2005.

[11] G. Loy, M. Eriksson, J. Sullivan, and S. Carlsson. Monocular 3d reconstruction of human motion in long action sequences. In *European Conference on Computer Vision*, 2004.

[12] K. Mikolajczyk, , R. Choudhury, and C. Schmid. Face detection in a video sequence – a temporal approach. In *Conference on Computer Vision and Pattern Recognition*, 2001.

[13] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering Human Body Configurations: Combining Segmentation and Recognition. In *Conference on Computer Vision and Pattern Recognition*, Washington, DC, 2004.

[14] D. Ormoneit, H. Sidenbladh, M. Black, and T. Hastie. Learning and tracking cyclic human motion. In *Neural Information Processing Systems*, pages 894–900, 2001.

Figure 8. Tracking results on another subject seen by a translating camera.

Figure 9. Robustness to misdetection. **First two rows:** Initial and refined poses for a sequence in which 3 consecutive key-poses are detected. **Last two rows:** Initial and refined poses for the same sequence when ignoring the central detection and using the other two. The initial poses are less accurate but the refined ones are indistinguishable.

[15] D. Ramanan, A. Forsyth, and A. Zisserman. Tracking People by Learning their Appearance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006. In press.

[16] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic Tracking of 3D human Figures using 2D Image Motion. In *European Conference on Computer Vision*, June 2000.

[17] H. Sidenbladh, M. J. Black, and L. Sigal. Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In *European Conference on Computer Vision*, 2002.

[18] G. Simon, A. Fitzgibbon, and A. Zisserman. Markerless tracking using planar structures in the scene. In *International Symposium on Mixed and Augmented Reality*, pages 120–128, October 2000.

[19] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3–D Human Motion Estimation. In *Conference on Computer Vision and Pattern Recognition*, San Diego, CA, June 2005.

[20] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *European Conference on Computer Vision*, 2002.

[21] L. Taycher, G. Shakhnarovich, D. Demirdjian, and T. Darrell. Conditional Random People: Tracking Humans with CRFs and Grid Filters. In *Conference on Computer Vision and Pattern Recognition*, 2006.

[22] A. Thayananthan, B. Stenger, P. Torr, and R. Cipolla. Tracking Articulated Hand Motion using a Kinematic Prior. In *British Machine Vision Conference*, pages 589–598, 2003.

[23] C. Tomasi, S. Petrov, and A. Sastry. 3d tracking = classification + interpolation. In *International Conference on Computer Vision*, pages 1441–1448, 2003.

[24] R. Urtasun, D. Fleet, and P. Fua. 3D People Tracking with Gaussian Process Dynamical Models. In *Conference on Computer Vision and Pattern Recognition*, New York, 2006.

[25] Q. Wang, G. Xu, and H. Ai. Learning object intrinsic structure for robust visual tracking. In *Conference on Computer Vision and Pattern Recognition*, Madison, WI, June 2003.

[26] Y. Wu, G. Hua, and T. Yu. Tracking articulated body by dynamic markov network. In *International Conference on Computer Vision*, 2003.