Monocular 3–D Tracking of the Golf Swing*

Raquel Urtasun
Computer Vision Laboratory
EPFL
1015 Lausanne, Switzerland
raquel.urtasun@epfl.ch

David J. Fleet
Dept. of Computer Science
University of Toronto
M5S 3H5, Canada
fleet@cs.toronto.edu

Pascal Fua
Computer Vision Laboratory
EPFL
1015 Lausanne, Switzerland
pascal.fua@epfl.ch

Abstract

We propose an approach to incorporating dynamic models into the human body tracking process that yields full 3–D reconstructions from monocular sequences. We formulate the tracking problem in terms of minimizing a differentiable criterion whose differential structure is rich enough for successful optimization using a simple hill-climbing approach as opposed to a multi-hypotheses probabilistic one. In other words, we avoid the computational complexity of multi-hypotheses algorithms while obtaining excellent results under challenging conditions.

To demonstrate this, we focus on monocular tracking of a golf swing from ordinary video. It involves both dealing with potentially very different swing styles, recovering arm motions that are perpendicular to the camera plane and handling strong self-occlusions.

1. Introduction

In spite of having received considerable attention, monocular tracking of human motion remains a difficult problem, especially in the presence of self-occlusions and movements perpendicular to the image plane. Even though early approaches used determinisitic optimization [6] for simple motions, recent advances involving more complex ones rely on multi-hypotheses optimization techniques [7, 9, 10, 12, 20] to resolve the ambiguities and to escape from the local-minima that are usually involved. They have been shown to be effective but entail ever increasing computational burdens as the number of degrees of freedom in the model increases.

In earlier work [22, 23], we have advocated the use of temporal motion models based on Principal Component Analysis (PCA) and inspired by those proposed in [18, 19, 24] to formulate the tracking problem as one of minimizing differentiable objective functions when using stereo data. The differential structure of these objective functions is rich enough to take advantage of standard hill-climbing optimization methods, whose computational requirements are much smaller than those of multiple-hypotheses ones and can nevertheless yield very good results.

Here, as shown in Figs. 1, 8, and 9, we extend this approach to monocular tracking, and demonstrate its ability to track such a complex fully 3–D motion as a golf swing. Unlike some recent approaches to incorporating dynamic models in 2–D [2], we recover full 3–D from a single fixed camera. This is important for golf because, at the top of the swing, the arm motion perpendicular to the camera plane is both large and very significant.

Of course, it could be argued that by using a strong motion model, we constrain the problem to the point where it becomes almost trivial. We will show that this is not the case and that our model still has sufficient flexibility not only to model very different golf swings, such as those of Figs. 1 and 9, but also to produce totally meaningless results if the image data is not properly exploited. In other words, our implementation embodies a happy middle ground between an over-constrained model that is too inflexible and one that is so loose that it makes the optimization very difficult. In our earlier work [23], we have found walking, running, and jumping, to be amenable to the kind of modeling we use here. We therefore believe our approach to be applicable not only to golf but also to many other motions that involve predictable movements.

In the remainder of this paper we first discuss related approaches and introduce our deterministic motion model. We then show how we use it to incorporate the kind of image information that can actually be extracted from video sequences acquired on golf courses under uncontrolled circumstances. Finally, we discuss our results in more detail.

2. Related Work

Modeling the human body and its motion is of enormous interest in the Computer Vision community, as attested by

*This work was supported in part by the Swiss National Science Foundation.

Figure 1: Tracking a full swing in a 45 frame sequence. First two rows: The skeleton of the recovered 3–D model is projected into a representative subset of images. Middle two rows: Volumetric primitives of the recovered 3–D model projected into the same views. Bottom two rows: Volumetric primitives of the 3–D model as seen from above.

recent surveys [15, 16]. However, existing techniques 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 3–D body pose is only partially recoverable from its projection in one single image. Reliable 3–D motion analysis therefore requires reliable 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 [1, 11, 17, 21] but requires very large sets of examples to be effective.
- **Track:** This involves predicting the pose in a frame given observation of the previous one. This 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 [7, 9, 10, 12, 20] or by using strong dynamic motion models as priors [2, 18, 19].

Neither technique is proven to be superior, and both are actively investigated. However, the tracking approach is the most natural one to use when a person is known *a priori* to be performing a given activity, such as walking, running, or jumping. Introducing a motion model becomes an effective means to constrain the search and increase robustness. Furthermore, instead of a separate pose in each frame, the output are the parameters of the motion model, that are physically usefull in subsequent tasks such as sport training, physiotherapy, diagnostics or recognition.

Models that represent motion vectors as linear sums of principal components are of particular interest to us and have been used effectively to produce realistic computer animations [3, 4, 5]. The PCA components are computed by capturing as many people as possible performing a specific

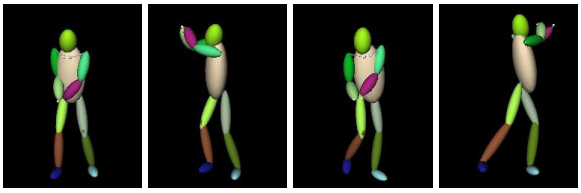

Figure 2: Key postures: Beginning of upswing, end of upswing, ball hit, and end of downswing.

activity, for example by means of an optical motion capture system, representing each motion as a temporally quantized vector of joint angles, and performing a Principal Component Analysis on the resulting set of vectors.

This representation has already been successfully used in our community [18, 19, 24], but almost always in a statistical context and without exploiting the fact that this parameterization allows the formulation of the tracking problem as one of minimizing differentiable objective functions, which allows for a lower computational complexity.

3. Motion Model

We represent the body of the golfer as a set of volumetric primitives attached to an articulated 3-D skeleton, as shown in the bottom rows of Fig. 1. Its pose is given by the position and orientation of its root node, defined at the level of the sacroiliac, and a set of joint angles. To build a motion model, we used the ten golf swing motions of the CMU database [8]. We identified the 4 key postures depicted by Fig. 2 in each motion and time warped the swings so that the key postures are all reached at the same time. We then sampled them at regular time intervals using quaternion spherical interpolation so that each swing can be treated as $N = 200$ samples of a motion starting at normalized time 0 and ending at normalized time 1.

In the models used here there are $D = 72$ degrees of freedom in addition to the 3D orientation and 3D position of the root. A swing is then represented by an *angular motion vector* Ψ of size $N * D = 14400$. Ψ is a column vector of the form:

$$\Psi = [\psi_{\mu_1}, \cdots, \psi_{\mu_N}]^T \tag{1}$$

where the ψ_{μ_i} are row vectors representing the joint angles at normalized time μ_i. The posture at a given time $0 \leq \mu_t < 1$ is estimated by interpolating the values of the ψ_{μ_i} corresponding to postures immediately before and after μ_t.

Assuming our set of training motions is representative of the motion we want to model, a motion vector Ψ can be approximated as a weighted sum of the mean motion Θ_0 and the first few principal directions of the training set, Θ_i, as follows:

$$\Psi \approx \Theta_0 + \sum_{i=1}^{m} \alpha_i \Theta_i \ , \tag{2}$$

where the α_i are scalar coefficients that characterize the motion, $m \leq M$ controls the fraction of the total variance of the training data that is captured in the subspace model, and $M = 10$ is the number of examples. For the small database we use, with $m = 4$ we were able to satisfactorily track golf swings. As will be discussed in Section 5.2, the database is clearly too small and we plan to augment it. However, based on previous experience with walking, running and jumping [23], we do not expect the required value of m to grow dramatically for the specific purpose of modeling golf swings, or more generally constrained athletic motions.

As will be discussed in Section 4, our tracking is formulated as the least-squares minimization of an objective function F with respect to the motion model parameters α_i, μ_t and the global motion G_t of the skeleton's root node, that is not included in the motion model. This involves computing the Jacobian of F. Assuming that $\frac{\partial F}{\partial \theta_j}$ is differentiable, that is that the derivatives of F with respect to the individual joint angles θ_j exist, this can be done analytically using the chain rule [22].

4. Least Squares Framework

Given that we operate outdoors in an uncontrolled environment and want to track golfers who are wearing their normal clothes, we cannot rely on any one image clue to give us all the information we need. Instead, we take advantage of several sources of information, none of which is perfect, but that together have proved sufficient for our purposes.

More specifically, we sequentially fit our motion model over sliding groups of f frames. For such a set of f frames, we take the state vector $\mathbf{S}$, to be

$$\mathbf{S} = [\alpha_1, \ldots, \alpha_m, \mu_1, \ldots, \mu_f, G_1, \ldots, G_f] \ , \tag{3}$$

where the α_i are the PCA weights of Section 3 common to the set of f frames, the μ_t are the normalized time associated to each frame, and the G_t represent the corresponding absolute position and orientation of the root node that vary in every frame.

We use the image data to write observation equations of the form $Obs(\mathbf{x}_i, \mathbf{S}) = \epsilon_i \ , \ 1 \leq i \leq n_{\text{obs}}$, where $\mathbf{x}_i$ is an observation, Obs a differentiable function whose value is zero for the correct value of $\mathbf{S}$ and completely noise free data, and ϵ_i is an error term. We then minimize a weighted sum of the squares of the ϵ_i residuals. Our system must be able to deal with observations coming from different sources. We therefore associate to each data point $\mathbf{x}_i$ an observation type $type_i$ and to each type a weight w^{type}. The weights are chosen so that the contribution of the different terms become commensurate in terms of their derivatives. Because the image data is noisy, we add a regularization term that forces the motion to remain smooth.

Figure 3: Poor quality foreground binary mask extracted from the images of Fig. 8.

The total energy F that we minimize therefore becomes

$$\sum_{i=1}^{nobs} w^{type_i} \left\| Obs^{type_i}(\mathbf{x}_i, \mathbf{S}) \right\|^2 \quad + \quad w_G \left\| G_t - \hat{G}_t \right\|^2 \quad (4)$$

$$+ w_\mu (\mu_t - \hat{\mu_t})^2 \quad + \quad w_\alpha \sum_{i=1}^{m} (\alpha_i - \hat{\alpha_i})^2 \quad ,$$

where Obs^{type} is the function that corresponds to a particular observation type, $\hat{G}^t$ and $\hat{\mu_t}$ are predicted values for the position and orientation of the root node and the predicted normalized time, and w_G, w_μ and w_α are scalar weights. We take $\hat{G}^t$ to be $G^{t-1}+\Delta G^{t-1}$ and $\hat{\mu_t}$ to be $\mu_{t-1}+\Delta\mu_{t-1}$, where ΔG^{t-1},$\Delta\mu_{t-1}$are the speeds observed in the previous set of frames.

We now turn to the description of the Obs^{type} functions for the data types we use and conclude the section by describing their complementarity.

4.1. Foreground and Background

Given an image of the background without the golfer, we can extract rough binary masks of the foreground such as those of Fig. 3. Note that because the background is not truly static, they cannot be expected to be of very high quality. Nevertheless, they can be exploited as follows. We sample them and for each sample $\mathbf{x}$ we define a *Background/Foreground function* $Obs^{fg/bg}(\mathbf{x}_i, \mathbf{S})$ that is 0 if the line of sight defined by $\mathbf{x}$ intersects the model and is equal to the distance of the model to the line of sight otherwise. In other words, $Obs^{fg/bg}$ is a differentiable function that introduces a penalty for each point in the foreground binary mask that does *not* backproject to the model. That penalty increases with the 3–D distance of the model to the corresponding line of sight.

Minimizing $Obs^{fg/bg}$ in the least squares sense tends to maximize the overlap between the model's projection and the foreground binary mask. This prevents the pose estimates from drifting, potentially resulting in the model eventually projecting at the wrong place and tracking failure.

4.2. Projection Constraints

To further constrain the location of six key joints—knees, ankles and wrists —and the head, we track their approximate image projections across the sequence.

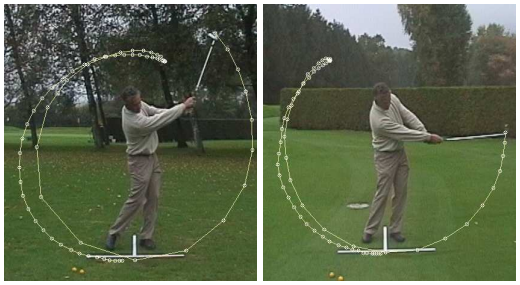

Figure 5: Detected club trajectories for the driving swing of Fig. 1 and the approach swing of Fig. 8. Note that one trajectory is much more extended than the other.

As shown in Fig. 4, for the ankles, knees and head, we use the WLS tracker [13] to take advantage of the slow dynamics of changes in image patches. WSL is a robust, motion-based 2–D tracker that maintains an online adaptive appearance model. The model adapts to slowly changing image appearance with a natural measure of the temporal stability of the underlying image structure. By identifying stable properties of appearance the tracker can weight them more heavily for motion estimation, while less stable properties can be proportionately down-weighted.

For the wrists, because the hand tends to rotate during the motion, we have found it more effective to use a club tracking algorithm [14] that takes advantage of the information provided by the whole shaft. It is depicted by Fig. 5 and does not require any manual initialization. It is also very robust to mis-detections and false alarms and has been validated on many sequences. Hypotheses on the position are first generated by detecting pairs of close parallel segments in the frames, and then robustly fitting a 2D motion model over several frames simultaneously. From the recovered club motion, we can infer 2–D hand trajectories.

For joint j, we therefore obtain approximate 2–D positions $\mathbf{x^j}_t$ in each frame. Given that the joint's 3–D position and therefore its projection are a function of S, we simply take the corresponding *joint projection function* $Obs^{joint}(\mathbf{x^j}_t, \mathbf{S})$ to be the 2–D Euclidean distance between the joint projection and its estimated 2-D location.

4.3. Point Correspondences

We use 2–D point correspondences in pairs of consecutive images as an additional source of information: We project the 3–D model into the first image of the pair, sample the projection, and establish correspondences for those samples in the second one using a simple correlation-based algorithm. Given a couple $\mathbf{x}_i = (p_i^1, p_i^2)$ of corresponding points found in this manner , we define a *correspondence function* $Obs^{corr}(\mathbf{x}_i, \mathbf{S})$ as follows: We backproject p_i^1 to the 3–D model surface and reproject it to the second image. We then take $Obs^{corr}(\mathbf{x}_i, \mathbf{S})$ to be the Euclidean distance in the image plane between this reprojection and p_i^2.

Figure 4: 2-D tracking of the ankles, knees, and vc2, using the WSL appearance-based tracker.

Figure 6: Tracking using only joint constraints vs using the complete objective function. Top: Using only joint constraints the problem is underconstrained and a multiple set of solutions are possible. Bottom: The set of solutions is reduced using correspondences.

4.4. Complementarity of the Data Terms

The projection observations of Section 4.2 more precisely constrain the projections of the ankles, knees, wrists and head. However, as shown in the top row of Fig. 6, these projection constraints are not sufficient by themselves. The correspondences of Section 4.3 are required to fully constrain the motion of both the lower and upper body. Of course, the correspondences by themselves would not be enough either: They are too noisy to be used alone because the golfer is wearing untextured clothing and the wrinkles produce correspondence motion that does not necessarily follow the golfer's true motion. As discussed above the foreground/background observations of Section 4.1 stop the estimates from drifting by guaranteeing that the model keeps on projecting roughly at the right place.

The example of Fig. 6 is significant because it shows that the model has sufficient flexibility to do the *wrong* thing given insufficient image data. In other words, even though we use a motion model, the problem is not so constrained that we are guaranteed to get valid postures or motions without using the images correctly.

5. Tracking

In this section, we first discuss the initialization of our tracking procedure, which only requires a minimal amount of manual intervention. We then present our results in more detail.

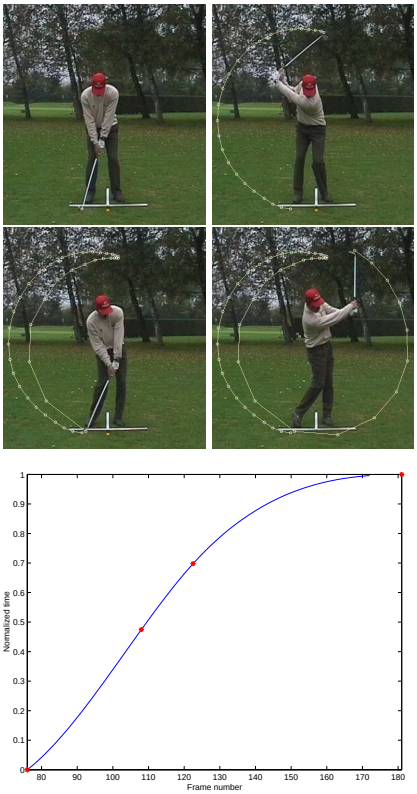

Figure 7: Assigning normalized times to the frames of Fig. 1. Fist two rows: We use the automatically detected club positions to identify the key postures of Fig. 2. Bottom row: The corresponding normalized times are denoted by red dots. Spline interpolation is then used to initialize μ_t for all other frames in the sequence.

5.1. Initialization

For each sequence, we first run the golf club tracker [14] discussed in Section 4.2. As shown in the first row of Fig. 7, the detected club positions let us initialize the μ_t parameters by telling us in which four frames the key postures of Fig 2 can be observed. As discussed in Section 3, the corresponding normalized times were defined when creating the database. We can therefore assign a normalized time

to all other frames in the sequence by spline interpolation, as shown in the bottom row of Fig. 7. As not everybody performs the motion at the same speed, this time is only a guess and will be refined during the actual optimization. This temporal alignment does not need to be precise, but it gives a rough initialization for each frame. This will not limit the application to other type of motions, since detection techniques [1, 11, 17, 21] has been proved effective for walking, running and could be easily extended to other athletic activities.

We then roughly position the root node of the body so that it projects approximately at the right place in the first frame and specify in that first frame the locations of the five joints to be tracked by WSL [13]. Note that all of this only requires a few mouse clicks and could easily be automated using posture detection techniques, given the fact that the position at the beginning of a swing is completely stereotyped. We can now start the tracking algorithm by setting all the PCA weights to zero and minimizing in a fully automated fashion the criterion of Eq. 5 three frames at a time.

5.2. Results

Figs. 1 and 8 depict complete driving swings performed by two different subjects whose motions were *not* recorded in the CMU database [8]. In both cases, we show projections of the recovered 3D model in a representative subset of the images. In Fig. 1, we also display the recovered 3D model, first projected in the original view and then as seen from above. Note the quality of the tracking in spite of the facts that the golfers are wearing relatively untextured clothing, their sizes are unknown and the cameras uncalibrated. To perform our computation, we used rough estimates of both the subjects size and the cameras focal length. In practice, this information could be made available to the system, thereby simplifying its task.

Fig 9 depicts a much shorter *approach* swing, where the club does not go as high as in a full swing, as evidenced by the very different club trajectories of Fig. 5. This is challenging for our system because the CMU database only contains driving swings. Our model nevertheless has sufficient flexibility to generalize to this new motion. Note, however, that the right leg bends too much at the end of the motion, which is a reflection of the small size of the database and of the fact that all the exemplars in it bend their legs in this particular fashion. One possible way to avoid this problem in the future is to use a larger database containing a great variety of training motions.

6. Conclusion

We have presented an approach to incorporating strong motion models that yields full 3–D reconstructions from monocular sequences using a single-hypothesis hill-climbing approach. This results in much lower computational complexity than current multi-hypotheses techniques. We have demonstrated it for monocular tracking of a golf swing from ordinary videos, which involves dealing with potentially very different swing styles, recovering arm motions that are perpendicular to the camera plane, and handling strong self-occlusions. The major limitation of the current implementation stems from the small size of the motion database we used, which we will remedy in the coming months.

We have obviously placed ourselves in a relatively constrained context, which is nevertheless far from simple and makes sense in terms of potential industrial applications. Furthermore, we believe there is also ample scope for broadening this approach given a ”library” of models such as the ones we have used here or those we developed in our earlier walking, running and jumping work [22, 23]: In a broader context, with specific motion models, we have traded the complexity of tracking for the complexity of knowing which model to apply. This might mean keeping several models active at any one time and selecting the one that fits best. This brings us back to multiple hypotheses tracking, but the multiple hypotheses are over models and not states. This might be much more effective than what many particle filters do because it ensures that the multiple hypotheses are sufficiently different to be worth exploring.

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*, pages III 54–65, Prague, May 2004.

[3] M. Alexa and W. Mueller. Representing animations by principal components. In *Eurographics*, volume 19, 2000.

[4] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating Faces in Images and Video. In *Eurographics*, Granada, Spain, September 2003.

[5] M. Brand and A. Hertzmann. Style Machines. *Computer Graphics, SIGGRAPH Proceedings*, pages 183–192, July 2000.

[6] C. Bregler and J. Malik. Tracking People with Twists and Exponential Maps. In *Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, June 1998.

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

[8] CMU database. http://mocap.cs.cmu.edu/.

[9] 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.

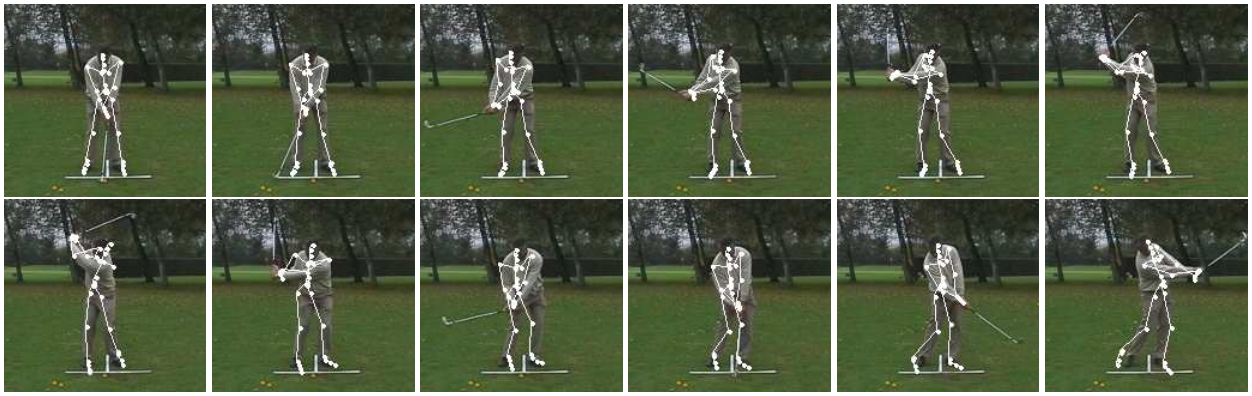

Figure 8: Tracking a 68 frame swing sequence. The skeleton of the recovered 3–D model is projected onto the images.

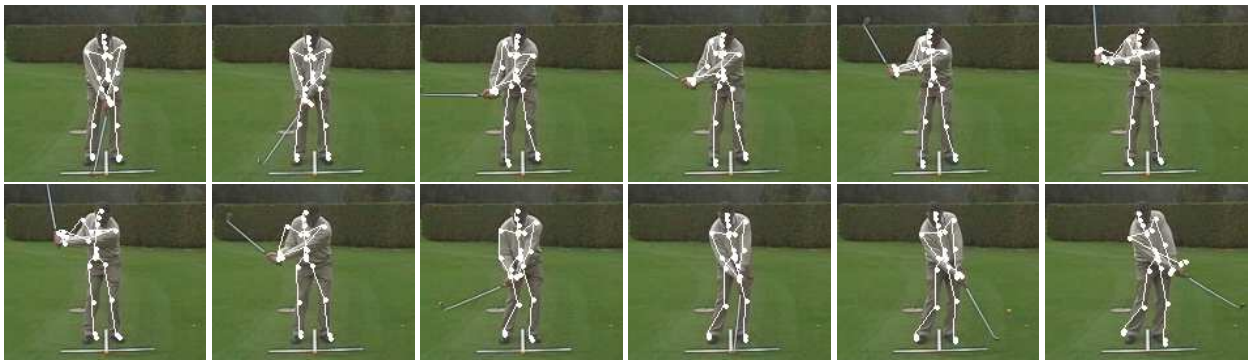

Figure 9: Tracking an approach swing during which the club goes much less high than in a driving swing. The skeleton of the recovered 3–D model is projected onto the images.

[10] 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.

[11] A. Elgammal and C. Lee. Inferring 3D Body Pose from Silhouettes using Activity Manifold Learning. In *CVPR*, Washington, DC, June 2004.

[12] M. Isard. and A. Blake. CONDENSATION - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, August 1998.

[13] A. Jepson, D. J. Fleet, and T. El-Maraghi. Robust on-line appearance models for vision tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1296–1311, 2003.

[14] V. Lepetit, A. Shahrokni, and P. Fua. Robust Data Association For Online Applications. In *Conference on Computer Vision and Pattern Recognition*, Madison, WI, June 2003.

[15] T. Moeslund. *Computer Vision-Based Motion Capture of Body Language*. PhD thesis, Aalborg University, Aalborg, Denmark, June 2003.

[16] T. Moeslund and E. Granum. A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding*, 81(3), March 2001.

[17] 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.

[18] D. Ormoneit, H. Sidenbladh, M. Black, and T. Hastie. Learning and tracking cyclic human motion. In *Advances in Neural Information Processing Systems 13*, pages 894–900, 2001.

[19] 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.

[20] C. Sminchisescu and B. Triggs. Kinematic Jump Processes for Monocular 3D Human Tracking. In *Conference on Computer Vision and Pattern Recognition*, volume I, Madison, WI, June 2003.

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

[22] R. Urtasun and P. Fua. 3D Human Body Tracking using Deterministic Temporal Motion Models. In *European Conference on Computer Vision*, Prague, Czech Republic, May 2004.

[23] R. Urtasun, P. Glardon, R. Boulic, D. Thalmann, and P. Fua. Style-based motion synthesis. *Computer Graphics Forum*, 23(4):799–812, December 2004.

[24] Y. Yacoob and M. J. Black. Parametric Modeling and Recognition of Activities. In *International Conference on Computer Vision*, pages 120–127, Mumbai, India, 1998.