BUNDLE ADJUSTMENT FOR MARKERLESS BODY TRACKING IN MONOCULAR VIDEO SEQUENCES

Ali Shahrokni, Vincent Lepetit, Pascal Fua
Computer Vision Lab, Swiss Federal Institute of Technology (EPFL)
ali.shahrokni,vincent.lepetit,pascal.fua@epfl.ch

KEY WORDS: Vision, Bundle Adjustment, Monocular Body Tracking

ABSTRACT

In recent years, because cameras have become inexpensive and ever more prevalent, there has been increasing interest in modeling human shape and motion from monocular video streams. This, however, is an inherently difficult task, both because the body is very complex and because, without markers or targets, the data that can be extracted from images is often incomplete, noisy and ambiguous. For example, correspondence-based techniques are error-prone for this kind of application and tend to produce many false matches.
In this paper, we discuss the use of bundle-adjustment techniques to address theses issues, and, more specifically, we demonstrate our ability to track 3D body motion from monocular video sequences. In earlier work, we have developed a robust method for rigid object monocular tracking and modeling. It relies on regularly sampling the 3-D model, projecting and tracking the samples in video sequences, and adjusting the motion and shape parameters to minimize a reprojection error. Here, we extend this approach to tracking the whole body represented by an articulated model. We introduce the appropriate degrees of freedom for all the relevant limbs and solve the resulting optimization problem. This scheme does not require a very precise initialization and we demonstrate its validity using both synthetic data and real sequences of a moving subject captured using a single static video camera.

1 INTRODUCTION

Observing the human body in motion is key to a large number of activities and applications such as security, character animation, virtual reality, human-machine interfaces, biomechanics studies, signaling in noisy environments, camera control, traffic and customer monitoring. All of the commercially available techniques for motion capture require either employing dedicated human operators or using ad-hoc sensors. This tends to make them:

- Cumbersome. The user needs to wear markers or other ad-hoc equipment which may be impractical, uncomfortable, constrain the user to a limited work space, be difficult to transport.
- Expensive. They require both hardware and skilled human operators.
- Slow: The data only becomes available after a lag required to process batches of images using manual techniques.

If motion capture could be become both automated and non-invasive, these limitations would disappear and many more applications would become practical. Multi-camera approaches (Delamarre and Faugeras, 1999, Gavrila and Davis, 1996, Plaenkers and Fua, n.d.) have the potential to achieve this goal**.** However, single camera solutions would be even more widely applicable. This is a challenging problem because it involves tackling such difficult issues as ambiguities associated with articulated motion seen from a single camera, very high dimensional search spaces, self-occlusions, and poor quality of image features in the absence of markers.

Many different strategies have been considered to handle the inherent difficulties of monocular body motion tracking. Some methods attempt to fit only a 2D model to the image stream (Morris and Rehg, 1998). This solution avoids the ambiguities and reduces the search space size but is inadequate for true 3D motion acquisition or analysis. A large majority of the fully 3–D methods rely on particle set-based tracker also known as the Condensation algorithm (Isard and Blake, 1998), sometimes in conjunction with a motion model. Some papers focus on improving to the original Condensation algorithm to perform the high-dimensional search (Sminchisescu and Triggs, 2001, Cham and Rehg, 1999), but these schemes are generic and do not exploit the specificities of the human tracking problem. Some papers focus on developing motion models (Sidenbladh et al., 2000) from real motions captured by sensors. Such models are useful to constrain the tracking but limit the application field.

Here, we propose a bundle adjustment framework to address the mentioned problems. We use multiple frames obtained from a single camera to compute the 3-D pose in each frame of a moving subject. We track 2–D points within the body outline over several images. We then assume that they are the projections of 3–D points on an articulated body model and estimate the motion parameters of this model by minimizing the reprojection errors in the least squares sense. Our approach is an extension of earlier model-based bundle adjustment techniques (Fua, 2000, Shan et al., 2001) that were designed to recover shape and pose parameters of heads in monocular video sequences. In these earlier papers the head is assumed to be rigid, whereas, in this work, we must account for the articulated nature of the body. We derive accurate estimation of all the body parameters for each view, even starting from a relatively weak initialization, without requiring a motion model. Thus the main contribution of this paper is the reformulation of model-based bundle adjustment for tracking of articulated motion in a monocular sequence. We will demonstrate the validity of our approach using both synthetic data and real monocular sequences of moving subjects.

In the next section, we introduce our formulation of the articulated bundle adjustment problem. In the following sections, we first propose a theoretical analysis of our approach in a simplified case and then show results on real monocular motion sequences.

2 PROBLEM STATEMENT

Bundle adjustment is usually used to simultaneously reconstruct a 3D scene and to recover camera positions. It can be also used to track a moving object by computing object positions with respect to a stationary camera and simultaneously adjusting the model shape parameters for a better fit.

2.1 Objective function

In our case we want to recover the of the body in a sequence of frames obtained by a single static camera observing a moving subject. We model the body as a set of connected limbs where each limb is represented by a well-defined surface such as ellipsoid or cylinder as shown in Figure 1. Therefore the whole body can be parametrized with

- a set of shape parameters $S = \{S_j \mid j = 1, ..., L\}$, where $S_j = \{C_i \mid i = 1, ..., n\}$, L is the number of body parts (head, trunk, legs and arms), and
- a set of pose parameters Θ_f defining the angles of each part wrt the one it is attached to as well as 6 DOF for defining the global position of the body.

S parameters are fixed over the whole sequence while Θ_f parameters are used to define the position at frame f. That means there are $N_f n(\Theta_f) + n(S) - 1$ unknowns to solve for, where N_f is the number of frames and $n(.)$ is the number of elements in a set, and the -1 comes from the fact that we can reconstruct the scene up to a scale factor, we assume that we know the length of at least one body part. This is a reasonable assumption as we can derive an approximate size of the body parts from anthropometric data. Our observations are regularly sampled points on a reference image for which we have an approximate 3D object pose. The main goal is to use these observations and an initial guess for the 3D position of the body, to compute the pose parameters for all frames. This is done by minimizing the distance between the observations and intersection of corresponding rays, cast from points on the model surface, with the image plane. The model points are computed from image points in the reference frame and the estimated 3D pose in that frame as illustrated in Figure 2 and explained in detail below.

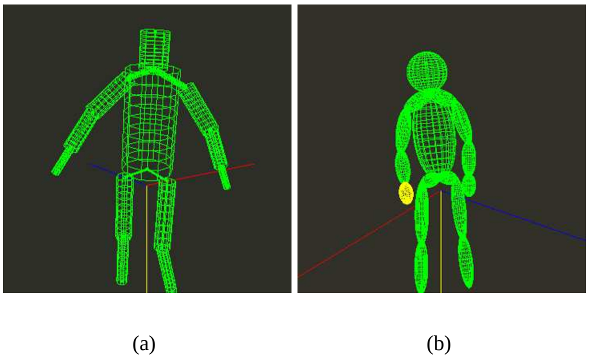

(a) (b)

Figure 1: Articulated models used to model a human body

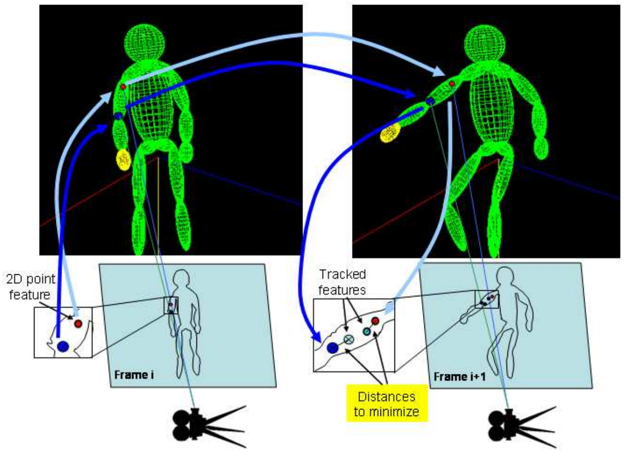


Figure 2: Computation of model distances from observations, for a given pose in a reference frame i

We have chosen to work on regularly sampled points rather than feature points to relax the constraints on the texture of the objects. We use a simple tracking algorithm based on mutual normalized image intensity cross correlation in two frames such as the one used in (Fua, 2000) to match 2D points. Each point on the reference image is tracked in zero or more frames in the sequence. We denote the tracked feature points as $p^l_{f,k}$, $f \in \{1, ..., N_f\}$ being the kth feature point on body part l tracked in the fth frame. It corresponds to an unknown 3D point P_k on the body part l as shown in Figure 3. For an image point on the reference image we compute the corresponding 3D point on the model and then we use its relative position on the object to compute the estimated position $\hat{p}^l_{f,k}$ of the tracked point $p^l_{f,k}$. The position of P$_k$ changes in each frame due to the motion of the body wrt the camera. More precisely the position of a fixed point on each body part depends on the orientation of the body limbs attached to limb l as well as the shape parameters of limb l: $P_k = P(\Theta_f, S_l)$. The projection of this point at frame f yields $\hat{p}^l_{f,k} = AP_k$, where A is the known camera projection matrix.

We want to minimize the reprojection error between the tracked points $p^l_{f,k}$,which might be corrupted with noise,

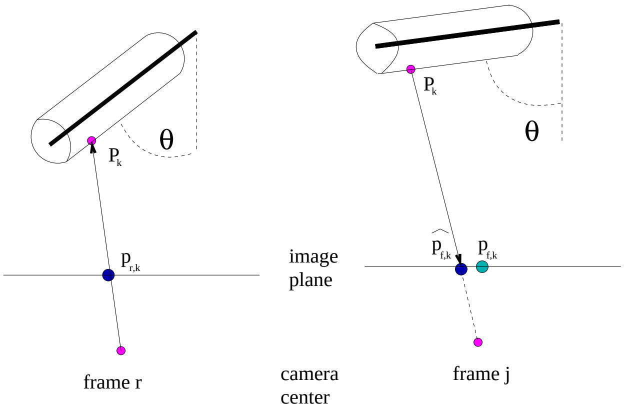


Figure 3: Reprojection error with transfer function relating image points in two frames by intersecting rays with the object surface.

and the corresponding projections $\hat{p}^l_{f,k}$

$$\min_{\{\Theta_f | f=1,\ldots,N_f\}} \sum_f \sum_l \sum_k \| \hat{p}^l_{f,k} - p^l_{f,k} \|^2 \qquad (1)$$

to obtain the pose parameters Θ_f for all frames.

To compute the 3D point P_k in frame f we first need to compute the position of the point on the limb l. This can be computed from the reference frame r for a given 3D position of the limb l and the projection of the point P_k, i.e. $\hat{p}^l_{r,k}$. Next we can compute the local position, in the reference system of limb l, of the point P_k by intersecting the line passing through $\hat{p}^l_{f,k}$ and the camera center with the surface of the shape representing limb l. Having the local coordinates we can compute the position of $P_k$in frame f and finally obtain $\hat{p}^l_{f,k}$ by projecting to the frame f. Therefore we can express the projection in frame f by a transfer function which takes a 2D point $\hat{p}^l_{r,k}$ in reference frame to another 2D point $\hat{p}^l_{f,k}$ on frame f.

$$\hat{p}^l_{f,k} = F(\Theta_f, S_l, \hat{p}^l_{r,k}) \qquad (2)$$

Replacing $\hat{p}^l_{f,k}$ with (2) in (1) allows solving the pose parameters.

2.2 Initial Guess

In practice, the computed initial position is usually erroneous especially if it is obtained from a single image,for example by assuming an orthographic camera model. In our implementation which is based on (Taylor, 2000), we obtain an estimation of the 3D pose in reference frame by first specifying for each joint whether or not it is closer to the camera than its parent joint. This is done by visual inspection of the image. Then we use an orthographic camera model to estimate the 3D position of joints wrt the join closest to the camera. Next we transform the computed coordinates of the joints into the camera reference system coordinates by solving a system of nonlinear equations. The results of initialization are not accurate and can confuse tracking methods for whose the position in each frame is a computed from the position in the reference frame. Using the bundle adjustment framework along with transfer function as defined in Equation 2, we can overcome this problem by including the pose parameters for the reference frame in the minimization formulation. This is described in detail in next section.

3 ANALYSIS OF SYNTHETIC ARTICULATION

In this section we discuss the applicability of our model-based bundle adjustment formulation to recovering poses for basic articulated structures. We begin by considering a simplified 2D problem. Using synthetic data we also show that the results carry over to 3D.

3.1 The double pendulum

For simplicity, here we will consider the simplified articulated structure shown in Figure 4 and that we refer to as a double pendulum. This is based on the fact that the general motion of human body can be decomposed into the motion of each limb (trunk, arms and legs). Considering the motion of each limb separately would facilitate further investigation of the optimization scheme in practical situations such as conditions raised by starting from an incorrect initial position. The pendulum's state in plane can be parametrized by the pivot position and length and angles of the arms $\Theta_f = \{x, y, \theta_1, \theta_2\}$ as shown in the Figure 4. We consider that the length of the first arm is known, which corresponds to the fact that we fix a scale in our reconstruction.

3.2 Minimization

We consider optimization over N_f frames (including the one with initialization) with N_p tracked points in all frames. Here the shape parameters S_l's are actually scale values that define the position of point on the pendulum shaft. The number of unknowns for the pendulum would be $N_p + 4N_f$ and number of terms in objective function 1 is N_pN_f. Therefore, for the system to be solvable we need to have $N_pN_f \geq N_p + 4N_f$ or $N_p(N_f - 1) \geq 4N_f$. For example to recover the state of a pendulum in two frames we need to have at least 8 tracked points. This is interesting because starting from a wrong initial state we can compute the state vector in all frames including the initial one and this solution is a global minimum in the sense that the projection errors in all frames would be minimized.

This is illustrated in Figure 5 for our pendulum model. Observations (circles on the horizontal projection line) are generated by regularly spaced sampling on the projection of the original pendulum shaft (dashed line) on the reference frame and computing their 3D position on shaft in that frame and then back projecting them to the next frame using ground truth values for next frame, we than add Gaussian noise to the tracked positions. We start from an incorrect position (thin solid line) and compute the position of points on the shaft and comparing the projection distance with the tracked points. Minimization of this

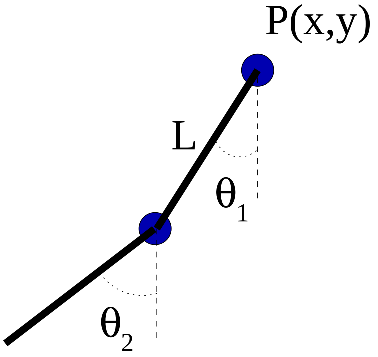


Figure 4: The double pendulum.

distance yields the correct pendulum position (thick solid line) in the two frames.

Due to large number of degrees of freedom in our bundle adjustment framework, starting from an initial state which is very far from the real one may mean falling into local minima which corresponds to incorrect poses. An example of this situation for our simplified double pendulum (8 DOF for two frames) model is illustrated in Figure 6. In order to avoid this problem we add a penalty term to the objective function which penalizes the poses for which the intersection of the casted ray from image points with the object surface induces shape parameters that are not acceptable. As an example, for the recovered pose in Figure 6 the intersection of casted rays with the pendulum makes one of the arms, for which we don't have a precise length, much longer than the original one. To solve this error we can make sure that the intersection length on the unknown arm lies on the acceptable range for arm length. The same argument applies to the general human model, for which we can use approximate lower and upper bounds for free shape parameters.

3.3 The 3-D case

In order to verify the application of our method human motion tracking we we ran a set of experiments similar to the ones in 2D for double pendulum on synthetically generated arm motion with two articulations in 3D space. There position and orientation of the joints are defined by $\Theta_f = \{x, y, z, \theta_{1x}, \theta_{1y}, \theta_{1z}, \theta_{2x}, \theta_{2y}, \theta_{2z}\}$, while the shape parameters are the parameters of the cylinders used to construct the model which are assumed fixed and known.

Observations are generated by adding noise to projection of points on the model. The noise model used was Gaussian with zero mean with random outliers to simulate the practical results of point matching in different frames. Figure 7 illustrates the model at frame 4 and its distance to observations before and after fitting and also the ground truth position which is similar to the recovered position in spite of high level of noise and outlier ratio.

4 EXPERIMENTAL RESULTS

Preliminary experiments were carried out on monocular sequences of human motion. To track the body motion in a sequence of frames, optimization is done over 3 frames with the frame in the middle as the reference image for which we have an initialization, and then the results of pose recovery for the frame following the reference frame are used to initial the next step using that frame as the reference image and so on. We used cylindrical shapes to model body parts as shown in Figure 1-a with appropriate DOF's. Figure 8 shows tracking results and point observations for a golf upswing. Because of occluded body position we use only half of the model for tracking. The tracking is lost during the course of upswing where the arm speed is fast. This is partly because of the low quality of the images and lack of high contrast in the images, but mainly due to the drift introduced by the 2D point matching algorithm. It is possible to improve the results if a more robust point matching algorithm such as the ones based on epipolar geometry is used. The other sequence consists of fast arm movements which is shown in Figure 9 with overlaid model and observations. We have chosen to track naked body because it is difficult to match points on skin. In spite of this problem the model follows the motion of the body for 20 frames and then loses track due to accumulated drift caused by failure in matching when the arm motion is fast.

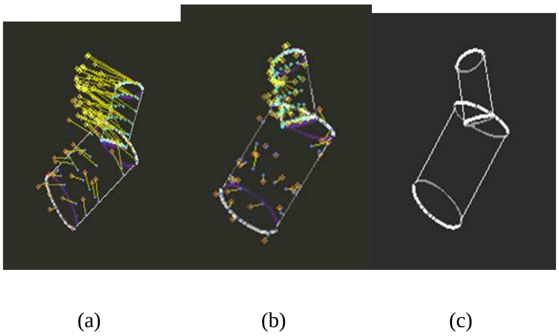
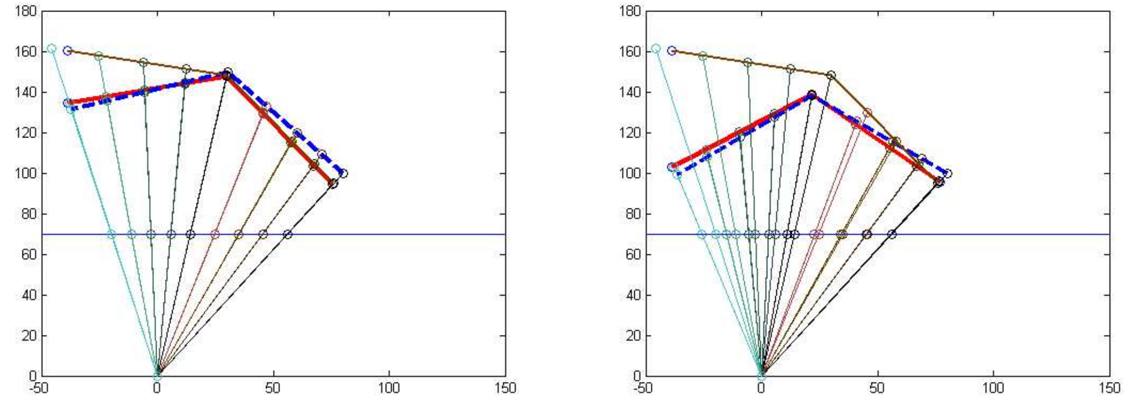


Figure 7: a) Initial position (from previous frame) and distances to noisy observations, b) final position and distances to noisy observations, and c) ground truth position used to generate observations.

5 CONCLUSION

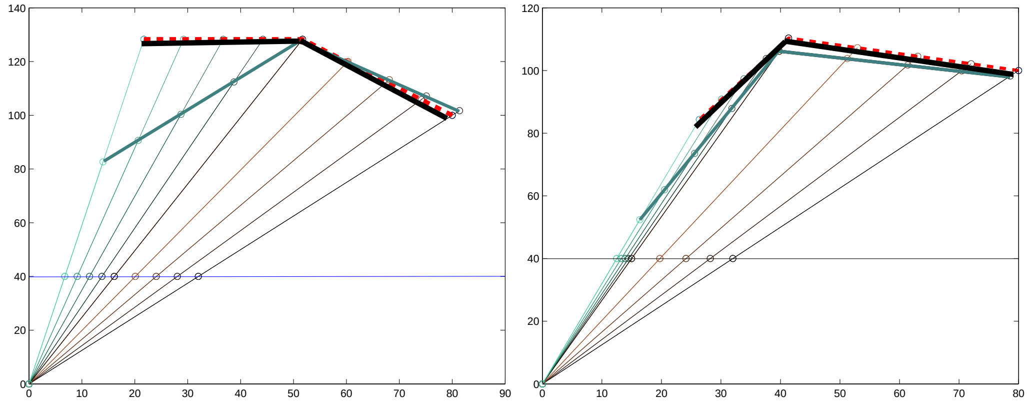
In this paper, we presented a model-based bundle adjustment framework for 3D tracking of articulated objects from monocular sequences. We introduced the appropriate degrees of freedom for all the relevant limbs and solved the resulting optimization problem. Having a set of tracked points in different frames without any 3D point correspondence in conjunction with an estimation of the 3D pose in a reference frame is enough to recover the pose in in the sequence of frames. Furthermore, it allows us to compensate for initialization error. In order to avoid local minima in the high dimensional pose space which tend to minimize the objective function by introducing drastic changes in unknown shape parameters, weconstrain limbs lengths to remain within anatomically feasible bounds. Our preliminary results on real and synthetic data demonstrate the validity of our approach. In the future we will incorporate a more robust point tracking kernel to avoid drift and to



(a)

(b)

Figure 5: Computing the position of pendulum in two frames (a is the reference frame) using noisy tracked points on a projection line. Dashed line is the ground truth, thin solid line is the initial position and thick solid line is the final position. The horizontal line is the projection line and the camera center is at origin. Circles on projection line are the observations.



(a)

(b)

Figure 6: Local minimum for pose parameters. Dashed line is the ground truth, thin solid line is the final position without using length constraints and thick solid line is the final position computed using length constraints. The horizontal line is the projection line and the camera center is at origin. Circles on projection line are the observations.

improve the results.

REFERENCES

Cham, T.-J. and Rehg, J., 1999. A multiple hypothesis approach to figure tracking. In CVPR Vol. 2, Ft. Collins, CO.

Delamarre, Q. and Faugeras, O., 1999. 3D Articulated Models and Multi-View Tracking with Silhouettes. In ICCV Corfu, Greece.

Fua, P., 2000. Regularized Bundle-Adjustment to Model Heads from Image Sequences without Calibration Data. In IJCV 38(2), pp. 153–171.

Gavrila, D. and Davis, L., 1996. 3d model-based tracking of humans in action : A multi-view approach. In CVPR San Francisco, CA.

Isard, M. and Blake, A., 1998. CONDENSATION - conditional density popagation for visual tracking. In IJCV 29(1), pp. 5–28.

Morris, D. and Rehg, J., 1998. Singularity Analysis for Articulated Object Tracking. In CVPR pp. 289–296.

Plaenkers, R. and Fua, P., n.d. Articulated Soft Objects for Multi-View Shape and Motion Capture. To appear in Pattern Analysis and Machine Intelligence.

Shan, Y., Liu, Z. and Zhang, Z., 2001. Model-Based Bundle Adjustment with Application to Face Modeling. In ICCV Vancouver, Canada.

Sidenbladh, H., Black, M. J. and Fleet, D. J., 2000. Stochastic Tracking of 3D Human Figures Using 2D Image Motion. In ECCV (2), pp. 702–718.

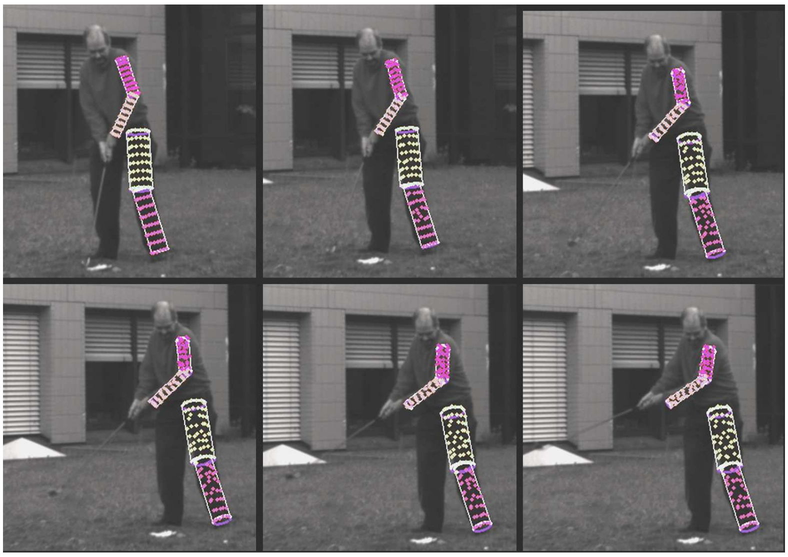


Figure 8: Tracking of a golf upswing in a low quality sequence

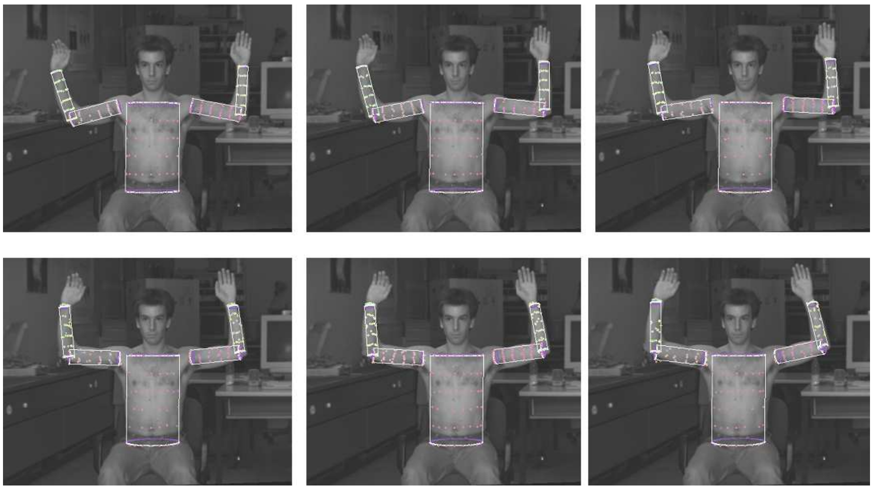


Figure 9: Tracking of naked body fast arm motion

Sminchisescu, C. and Triggs, B., 2001. In CVPR Covariance Scaled Sampling for Monocular 3D Body Tracking.

Taylor, C., 2000. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In CVIU 80(3), pp. 349–363.