Monocular 3D Reconstruction of Human Motion in Long Action Sequences

Gareth Loy, Martin Eriksson, Josephine Sullivan, and Stefan Carlsson

Computational Vision & Active Perception Laboratory (CVAP),
Department of Numerical Analysis and Computer Science,
Royal Institute of Technology (KTH), S-100 44 Stockholm, Sweden
{gareth,eriksson,sullivan,stefanc}@nada.kth.se

Abstract. A novel algorithm is presented for the 3D reconstruction of human action in long (> 30 second) monocular image sequences. A sequence is represented by a small set of automatically found representative keyframes. The skeletal joint positions are manually located in each keyframe and mapped to all other frames in the sequence. For each keyframe a 3D key pose is created, and interpolation between these 3D body poses, together with the incorporation of limb length and symmetry constraints, provides a smooth initial approximation of the 3D motion. This is then fitted to the image data to generate a realistic 3D reconstruction. The degree of manual input required is controlled by the diversity of the sequence's content. Sports' footage is ideally suited to this approach as it frequently contains a limited number of repeated actions. Our method is demonstrated on a long (36 second) sequence of a woman playing tennis filmed with a non-stationary camera. This sequence required manual initialisation on $< 1.5\%$ of the frames, and demonstrates that the system can deal with very rapid motion, severe self-occlusions, motion blur and clutter occurring over several concurrent frames. The monocular 3D reconstruction is verified by synthesising a view from the perspective of a 'ground truth' reference camera, and the result is seen to provide a qualitatively accurate 3D reconstruction of the motion.

1 Introduction

This paper addresses the challenge of generating a qualitatively accurate 3D reconstruction of the actions performed by an individual in a long ($\sim$30 second) monocular image sequence. It is assumed the individual is not wearing any special reflective markers or clothing. Any solution must be able to cope with the multitude of difficulties that may arise over several concurrent frames: severe self-occlusion, unreliability of methods for limb and joint detection, motion blur, and the inherent ambiguities in reconstructing rigid links from monocular images [15]. Until now, the only approach guaranteed to produce a complete and accurate reconstruction in such circumstances is: *for each frame in the sequence, manually locate the skeletal joints and perform 3D reconstruction using the method of* [15]. The latter involves solving the forward/backward binary ambiguity for each rigid link by inspection and estimating the relative lengths of each limb. For very short sequences this is a relatively painless procedure, but rapidly becomes impractical for longer sequences.

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3024, pp. 442–455, 2004.

The traditional tracking approach to human motion capture [7] is to perform manual initialisation at the beginning of the sequence and then update the estimate of the reconstruction over time in accordance with the incoming data. In contrast we consider the entire sequence and approximate the actions present by a set of representative frames (automatically determined from the sequence) and from these obtain a coarse description of the subject's motion. Finer detail is added by locating the skeletal joints in each frame by extrapolating from manually initialised joint locations on the representative frames.

The degree of manual input required is controlled by the diversity of the sequence's content. Sports' footage is ideally suited to this approach as it frequently contains a limited number of repeated actions. Throughout this paper the ideas and methods developed are illustrated and tested on a 36 second sequence of a woman playing tennis. Our results are verified by synthesising a view of the 3D reconstruction from the perspective of a reference camera not used for the reconstruction.

The motivation for pursuing this problem together with a review of related research is presented in section 2. An overview of the algorithm is given in section 3. Section 4 details the grouping performed to obtain a keyframe representation of a sequence. Building upon this representation, the skeletal joint locations in each frame are estimated (section 5). The procedure for constructing the 3D reconstruction of the sequence is given in section 6, and the final reconstructions achieved for the tennis sequence are displayed in section 7 prior to the concluding remarks.

2 Background

Markerless human motion capture has drawn growing interest in recent years. The majority of systems developed have used multiple cameras to capture the subject [2,3,7]. However, stereo systems are rare outside of research laboratories and studios, and the bulk of videos of human activity are monocular. This, together with the comparative ease of capturing monocular sequences, motivates the monocular problem as one of more than purely academic interest.

Several researchers have tackled the challenge of human motion capture from monocular sequences, and some impressive results have been achieved over short sequences [13, 10]. Sminchisescu and Triggs [12,13] have achieved the most successful results to date in monocular markerless 3D human motion capture. Their algorithms are based upon propagating a mixture of Gaussians pdf, representing the probable 3D configurations of a body over time. Success relies upon performing efficient and thorough global searches of the cost surface associating the image data to potential body configurations. These methods have proved effective on relatively short sequences. However, it is an open question, whether the propagation of a multi-modal distribution, without an explicit mechanism for re-initialisation, is sufficient for long sequences.

Potential disruptions to smooth tracking conditions can be bridged by imposing priors on the dynamics of the configuration of the body. These have been used to some effect [11,10,1]. However, this comes at a cost. The motions present in a novel sequence may not be adequately described by the priors in use, and the appropriate trade-off between fitting the image data and fulfilling the prior constraints has to to decided. Also for long sequences of diverse motion (e.g. tennis) no one dynamical model can fully explain the motions present, necessitating the introduction of some form of recognition.

The general problem with tracking long sequences is that it is difficult to encapsulate the diversity of motion in a prior model. However, it is possible to summarise the motion in such a sequence. Several researchers have summarised the content of video by detecting and describing the actions (or subjects) present [17,8] either by clustering together frames or sequences of frames with similar properties. Toyama and Blake [16] showed that actions in a sequence could be summarised by a set of keyframes (exemplars) extracted from the sequence, and preceded to describe a novel video clip as a sequence of warped versions of these keyframes. A similar approach has been taken in more recent work [14,6], where sophisticated methods are used to match hand-defined keyframes to individual frames. Furthermore, by identifying specific joint locations on each keyframe, it was possible to localise these joint positions throughout a sequence. These methods, though only applied to short sequences, show an approach to tracking driven by *pose recognition*. This circumvents the problem of initialisation and is resistant to complete failure due to tracking loss, thereby opening the way to track long sequences.

This paper extends the keyframe-based approach of Sullivan and Carlsson [14] to long sequences with no prior learning and no pre-defined keyframes. A subsequent 3D reconstruction is performed using the method of [15].

3 Overview of Algorithm

Figure 1 gives an overview of the algorithm developed in this paper, from the initial extraction of the keyframes summarising the sequence, through the labelling of skeletal joint positions, formation of 3D keyframes, and interpolation of 3D keyframes, to the final 3D reconstruction.

Automatically representing the sequence by a set of keyframes requires measuring the similarity between the poses present in every pair of frames in the sequence. A distance matrix summarises these similarities and is used as the basis for finding the representative poses which in turn are encapsulated in keyframes summarising the sequence. The second layer in figure 1 encompasses the initialisation of each keyframe: the 2D skeletal joints are manually labelled and their corresponding 3D reconstructions created [15]. The 2D skeletal joints are then automatically determined throughout the sequence using the 2D keyframes and the keyframe assignment for each frame [14,6]. This involves

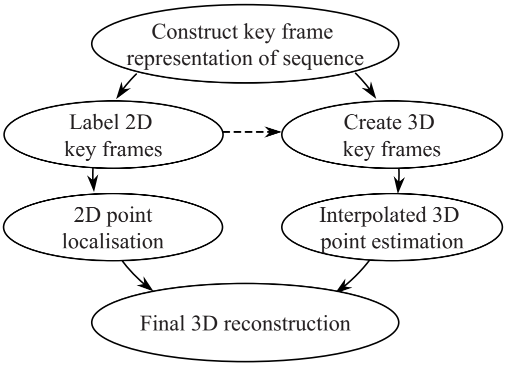


Fig. 1. Overview of the algorithm.

approximating the warp of the assigned keyframe to each frame and transferring the defined skeletal joints accordingly.

Next follows an initial estimation of a smooth 3D reconstruction of the sequence, whereby each frame deemed sufficiently close to a 3D keyframe is replaced by that keyframe. Interpolation occurs between these frames to estimate the intermediate frames. Finally the interpolated 3D reconstruction is refined to fit the estimated 2D joint locations throughout the sequence. This is achieved by minimising the reprojection error, while taking into account motion smoothness and imposing limb length and symmetry constraints. This ensures that any errors in the 2D data do not result in invalid reconstructions of the skeleton.

4 Defining Keyframes

We are interested in extracting, from a sequence $\mathcal{I} = \{1, \cdots, N\}$, a set of keyframes $\mathcal{K} \subset \mathcal{I}$ which span the body poses in $\mathcal{I}$. Besides providing a summary of the content of the sequence, each keyframe will assist in the skeletal joint localisation in frames of similar appearance. Such frames are considered *well-represented* by a keyframe. Thus $\mathcal{K}$ has an associated set $\mathcal{W}_{\mathcal{K}} \subset \mathcal{I}$ of frames it well-represents. The poses between two well-represented frames less than T frames apart, may be approximated by interpolating between the well-represented frames. These interpolatable frames define a set $\mathcal{J}_{\mathcal{K}} \subset \mathcal{I}$.

We wish to choose the least number of keyframes that enable an accurate description of the pose in a percent α of the sequence's frames. That is, we aim to find the $\mathcal{K}$ with minimal cardinality such that

$$|\mathcal{W}_{\mathcal{K}} \cup \mathcal{J}_{\mathcal{K}}| \geq \alpha N \tag{1}$$

Keyframe selection is based upon a distance matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$ describing the similarity in body pose between every pair of frames in the sequence. Below we explain how $\mathbf{D}$ is computed and then analysed to produce $\mathcal{K}$.

4.1 Measuring Pose Similarity between Frames

The subject is localised by finding the head and feet positions in each frame. This is done by sequentially applying colour histograms, low-pass filtering, and a radial symmetry operator [5] to detect round and elliptical regions of the appropriate scale and colour to correspond to either a head or foot. A plausible series of head and feet positions is isolated by finding the most temporally consistent path of the candidate locations through the sequence [4]. Based on the computed head and feet locations, a bounding box is estimated for the subject. Figure 2 illustrates this process.

Target regions of homogeneous colour are then extracted, and represented by *directed edge elements*: The edges of each region are sampled at regular intervals. Each sample point is represented by a point vector tangent to the edge and oriented so the interior of the target region is to its left, see figure 2(e).

Pairs of images can now be compared by computing a correspondence field between the edge points. The frames are aligned using the tracked head and feet locations, and each edge element matched to the closest edge element in the other image from the

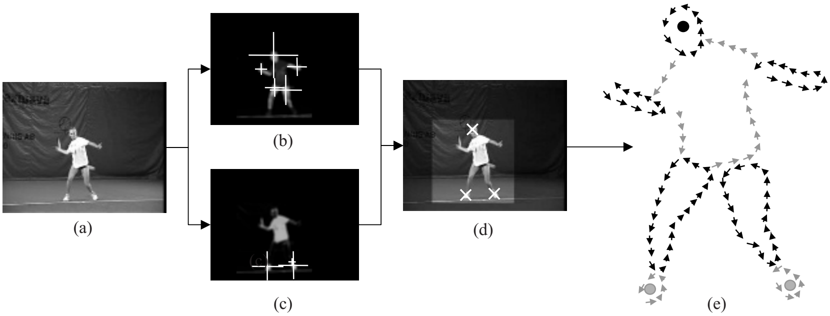


Fig. 2. (a) Original image, (b) head- and (c) feet-like colours highlighted, low-pass filtered and with peaks in radial symmetry indicated — the magnitude of each peak is shown by the size of the cross — (d) identified head and feet regions and resulting bounding box, (e) directed edge elements of target regions.

same coloured target regions, and whose orientation differs by less than 45 degrees. A comparison of the body poses can then be computed by considering the average distance between corresponding points, together with the percentage of edge elements for which a corresponding match was found.

4.2 Distance Matrix

Using the method described in section 4.1, we can determine the distance and the percentage of successfully matched points between every i^{th} and j^{th} frame in the sequence. Putting the respective output into the matrices $\mathbf{B}$, $\mathbf{A} \in \mathbb{R}^{N \times N}$, an initial distance matrix $\mathbf{C}$ is then computed by combining these as

$$\mathbf{C}(i,j) = \mathbf{A}(i,j)\mathbf{B}(i,j) + (1 - \mathbf{A}(i,j)) \max \mathbf{B}$$

The resulting matrix $\mathbf{C}$ gives a good indication of the dissimilar and similar frames. However, it can be improved. When the inter-frame distance is sufficiently small, $\mathbf{C}(i,j) < \beta$, frames i and j are extremely likely to contain the same pose. In this case the corresponding i^{th} and j^{th} rows (and columns) of $\mathbf{C}$ should be almost identical, and any observed differences can be treated as noise. The final distance matrix $\mathbf{D}$ is formed by replacing each row and column of $\mathbf{C}$ with the average of all the rows and columns corresponding to frames to which it has a distance less than β. This reduces the noise giving a cleaner distance matrix.

Figure 3 shows $\mathbf{D}$ for the upper body for an 1800 frame tennis sequence, with several example frames and their corresponding rows and columns in the matrix. The dark rectangular regions in the matrix correspond to periods where there is little change between frames. For the tennis sequence this equates to the player standing still in between strokes, such as in frames 448 and 1662. Dark diagonals (off the main diagonal) correspond to distinct repeated events, such as the forehand (614, 1258) and backhand (174, 1032) frames. Note that there is only one such dark diagonal in the rows and columns corresponding to frames 174 and 1032. This is because there are only two backhands in the sequence, and thus only one repeated event.

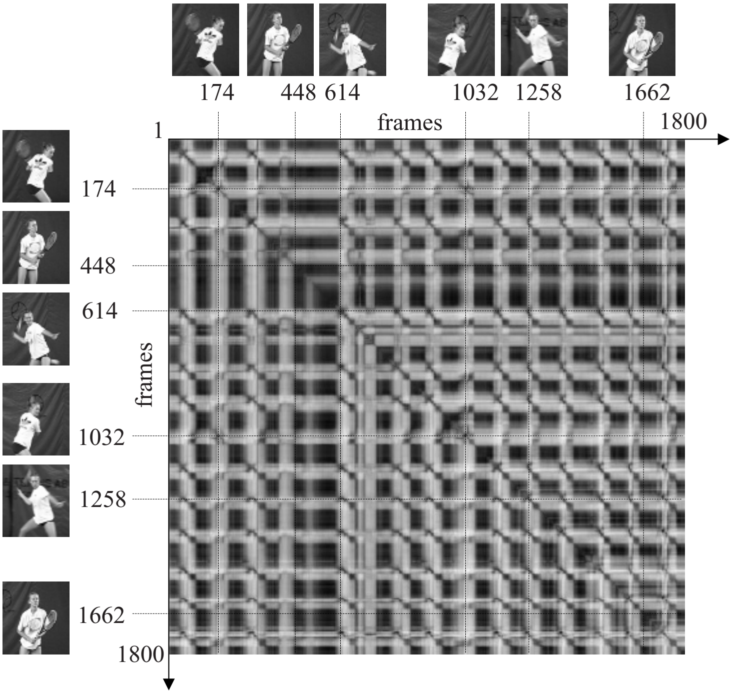


Fig. 3. The distance matrix $\mathbf{D}$ for the upper body pose over a 1800 frame (36 second) tennis sequence, with several sample frames. Short dark diagonals correspond to forehands and backhands, and dark rectangular regions indicate periods where the player is standing still.

4.3 Keyframe Selection

We define a criterion for considering one frame to be well-represented by another. Recall in section 4.2 that if $\mathbf{D}(i, j) < \beta$, then frame i and j are considered to exhibit the same pose. We say that such frames are *well-represented* by each other.

We now describe an algorithm to find a $\mathcal{K}$ with minimal $|\mathcal{K}|$ which fulfills equation (1). Keyframes are iteratively selected to minimize the average distance of all frames from their neighbouring well-represented frames. Firstly, define $C_{\mathcal{K}}$ as:

$$C_{\mathcal{K}} = \sum_{i=1}^{N} \left(\min_{f \in \mathcal{W}_{\mathcal{K}}, f<i} |i - f| + \min_{f \in \mathcal{W}_{\mathcal{K}}, f>i} |i - f| \right) \tag{2}$$

Then set $\mathcal{K}^0 = \emptyset$. Keyframes are repeatedly selected according to:

$$\mathcal{K}^{(t+1)} = \mathcal{K}^{(t)} \cup \{j\} \tag{3}$$

where

$$j = \operatorname*{arg\,min}_{1 \leq k \leq N, k \notin \mathcal{K}^{(t)}} C_{\mathcal{K}^{(t)} \cup \{k\}} \tag{4}$$

until the criterion in equation (1) is satisfied.

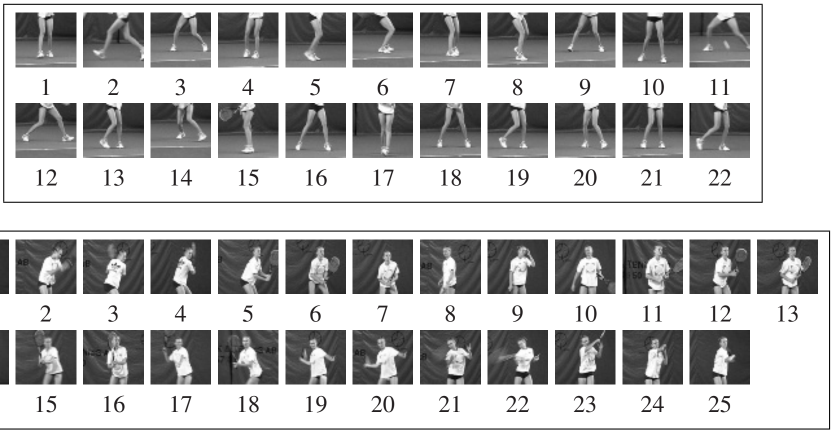


Fig. 4. The 22 lower and the 25 upper body keyframes extracted from the 36 second sequence.

This algorithm was applied to extract keyframes from an 1800 frame sequence of a woman playing tennis with T set to 10 and $\alpha = 0.95$. The upper and lower body were divided, and separate distance matrices and key frames determined for each. 25 key frames were required for the upper body and 22 for the lower body in order to satisfy equation (1) (see figure 4).

4.4 A Keyframe Representation of the Sequence

Figure 5 shows an example upper body and lower body keyframe and the associated well-represented frames from the sequence.

By representing each frame by its closest keyframe we can examine the occurrence of different body poses throughout the sequence. Figure 4 shows all the keyframes extracted from the sequence and figure 6 shows which keyframe best represents each frame throughout the sequence. This graph characterises the pose variation in the sequence and the forehands and backhands are easily identified respectively by the strong peaks and troughs in the graph in figure 6.

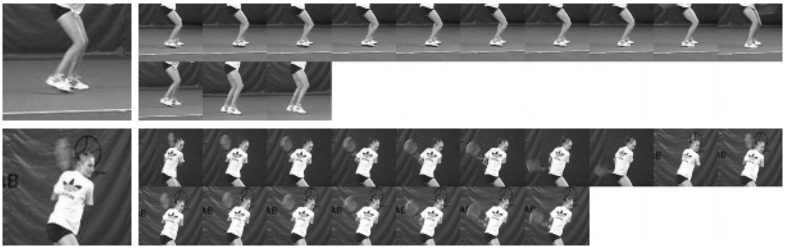


Fig. 5. Example upper and lower body keyframes, and the frames well-represented by these keyframes.

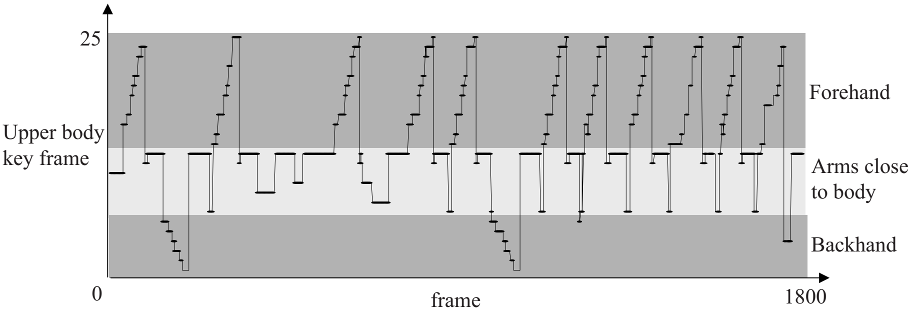


Fig. 6. Occurrence of frames associated with the various keyframes throughout sequence. This graph is for the upper body.

5 Locating Joint Positions

For each keyframe, $k \in \mathcal{K}$, its n skeletal joints, $\mathbf{x}_k = (x_{1,k}, \cdots, x_{n,k})$ $x_{i,k} \in \mathrm{I\!R}^2$, are manually annotated. Points from the appropriate keyframe are then automatically mapped to every frame in the sequence to obtain an estimate of $\mathbf{x}_{1:N} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)$. Figure 7 shows an annotated keyframe k, and joint locations estimated for a frame t, assigned to this keyframe. The aligned keyframe edges have been superimposed onto Figure 7(b). Each joint in the keyframe has associated edge points in its vicinity and the correspondences found between these edge points and the edge points in the frame t define a translation. This translation is used to transfer the joint from the keyframe to frame t. Once an estimate of each joint in frame t is obtained, it is refined using the appearance of the joints in the keyframe, and enforcing the apparent limb length ratios evident in the keyframe [14]. Figure 7(c) shows the final estimates.

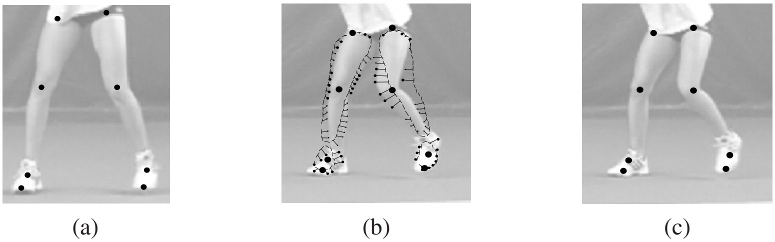

(a) (b) (c)

Fig. 7. (a) annotated keyframe k, (b) point correspondences between keyframe and well-represented frame, and (c) joint locations estimated for the well-represented frame t.

6 3D Reconstruction

The human skeleton can be modelled as an articulated chain with n_l links. Given the projection of the skeletal joint locations $\mathbf{x}_t$ onto the image plane, the number of qualitatively different reconstructions, $\mathbf{X}_t$, is bounded by 2^{n_l}[15] (assuming orthographic imaging), as each link can point either toward or away from the image plane. For an N

frame sequence, the number of possible reconstructions explodes to $2^{n_l N}$. This enormous search-space can be pruned by imposing the physiological limitations of the human body [13] and bounding the motion between adjacent frames. Without prior information, estimating the skeleton's configuration $\mathbf{X}_{1:N}$ over the sequence requires deciding the optimal binary labelling at each frame based on heuristic continuity measures.

Therefore, the crucial issue is the generation of prior information about the 3D configuration of the subject in the video. From the previous section we have a set of keyframes, $\mathcal{K}$, which span the 2D poses in the sequence. The 3D reconstruction of these keyframes provides an approximate basis for the 3D poses exhibited in the sequence. Thus with a limited amount of manual effort we have obtained some crucial priors. The next section describes how these 3D keyframes are used to create a smooth initial estimate, $\mathbf{X}^0_{1:N}$, of the 3D configuration of the subject throughout the sequence.

6.1 Establishing a Smooth Representative Reconstruction

The elements of $\mathcal{W}_{\mathcal{K}}$ and their corresponding keyframe assignments define the frames in the sequence that are well approximated by the 3D keyframes. Replacing each of these frames with its appropriate keyframe, and using these as control points in a spherical linear interpolation (slerp) process [9] allows the approximation of intermediary frames not in $\mathcal{W}_{\mathcal{K}}$. Keyframes have been chosen to ensure that the temporal distance interpolated is never large (equation (1)). However, frequently temporally adjacent frames in $\mathcal{W}_{\mathcal{K}}$ are assigned to the same keyframe. In reality they do not correspond to exactly the same 3D pose. One of the frames' 3D poses will, in general, match the keyframe more accurately than the others, and the other frames are better approximated by interpolation between the keyframes that temporally bound them.

To this end, temporal runs of frames in $\mathcal{W}_{\mathcal{K}}$ that are well-represented by the same keyframe are identified. The fit of each frame in the run to the 3D keyframe is ranked (ranking is based on a robust measure of the Euclidean distance between the reprojected 3D keyframe and the frame's estimated 2D joints). The lowest ranked frames in each run are iteratively omitted from the set of control points, subject to the criterion that T must be the maximum distance between control points.

Once the final control points have been decided, the interpolation is performed to obtain $\hat{\mathbf{X}}^0_{1:N}$. Figure 8 summarises the interpolation process.

6.2 Fitting the Smooth Motion Estimate to the Joint Data

The last task is to refine the 3D reconstruction by allowing the localised joint locations $\hat{\mathbf{x}}_{1:N}$ to influence $\hat{\mathbf{X}}^0_{1:N}$. However, the localised joint locations may contain outliers, be corrupted by noise and suffer from missing estimates due to self-occlusion. To ensure robustness to these factors, the final estimate of $\mathbf{X}_{1:N}$ is forced to be a valid trajectory of a human skeleton.

Define $\mathcal{M}_N$ as the manifold describing all valid trajectories of length N of the skeleton. Then:

$$\hat{\mathbf{X}}_{1:N} = \arg\min_{\mathbf{X}_{1:N}} E(\mathbf{X}_{1:N}) \quad \text{subject to} \quad \hat{\mathbf{X}}_{1:N} \in \mathcal{M}_N. \tag{5}$$

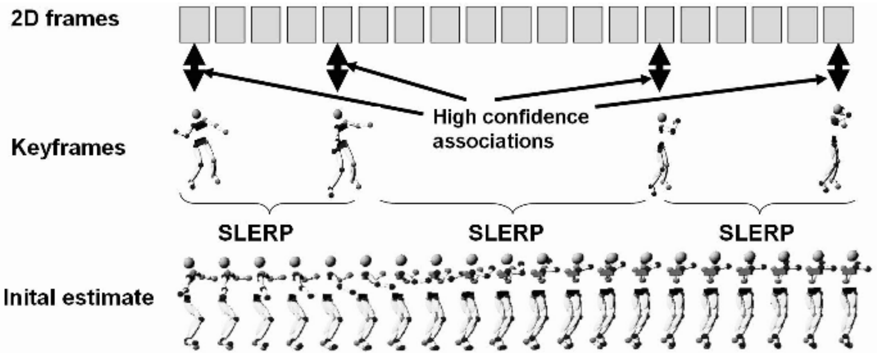


Fig. 8. Visualisation of the generation of a smooth and plausible trajectory of the 3D skeleton that approximates the content of the video.

where E is a cost function based on the sum of squared differences between $\hat{\mathbf{x}}_{1:N}$ and the orthographic projection of $\mathbf{X}_{1:N}$ (denoted by $\mathbf{x}'_{1:N}$):

$$E(\mathbf{X}_{1:N}) = \parallel \mathbf{x}'_{1:N} - \hat{\mathbf{x}}_{1:N} \parallel^2 \tag{6}$$

There is no easy characterisation of $\mathcal{M}_N$, so enforcing $\hat{\mathbf{X}}_{1:N}$ to belong to $\mathcal{M}_N$ is difficult. However, all members of $\mathcal{M}_N$ must exhibit constant limb-length throughout the sequence, and each joint trajectory must follow a smooth path. By forcing $\hat{\mathbf{X}}_{1:N}$ to satisfy these constraints, $\hat{\mathbf{X}}_{1:N}$ will be on or close to $\mathcal{M}_N$.

```
Step 1: Translate, rotate and scale $\hat{\mathbf{X}}^0_{1:N}$ to fit the 2D data
Step 2: Set $i = 1$.
Step 3: Gradient descent:
        $\hat{\mathbf{X}}^i_{1:N} = \hat{\mathbf{X}}^{i-1}_{1:N} - \lambda \nabla_{\mathbf{X}_{1:N}} E|_{\hat{\mathbf{X}}^{i-1}_{1:N}}, \quad 0 < \lambda \leq 1.$
Step 4: Enforce constraints: $\hat{\mathbf{X}}^i_{1:N} \in \mathcal{M}_N$.
Step 5: Increment $i$ by one and goto Step 3. (until convergence)
```

Fig. 9. The iteration steps involved in finding $\hat{\mathbf{X}}_{1:N}$.

By construction $\hat{\mathbf{X}}^0_{1:N} \in \mathcal{M}_N$. Therefore, it is used as the initial guess for the solution of the minimisation problem posed in equation (5). Figure 9 gives an outline of how the minimisation proceeds. At the end of each iteration, enforcing $\hat{\mathbf{X}}^i_{1:N} \in \mathcal{M}_N$ is approximated by resetting the limb-lengths to their correct value, and applying a low-pass filter to the trajectories of each joint. A large λ yields faster convergence, but makes it more difficult to re-project the solution back onto $\mathcal{M}_N$. We used $\lambda = 0.2$ for our experiments.

Figure 10 shows how a 3D keyframe is refined in 3D to match the image data. Here the same 3D keyframe is modified to form two different 3D reconstructions to match two different forehand frames, capturing the subtle differences between the two forehand strokes.

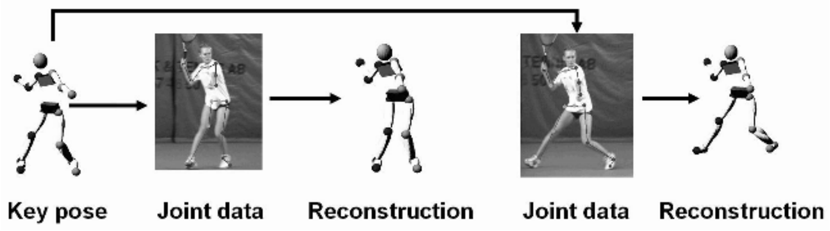


Fig. 10. Result of refining a key-pose based on image-data. The key-pose is refined according to the different images, resulting in two different 3D poses.

7 Results

Our algorithm was applied to reconstructing a 36 second tennis sequence filmed with a non-stationary camera. During the sequence the player moves about the baseline and plays several forehand and backhand strokes. Our results were verified by synthesising a view of the 3D reconstruction from the perspective of a reference camera. Figure 11 shows the experimental setup together with 3D reconstructions throughout the sequence and associated 'ground-truth' frames from the reference camera. Figure 12 shows a reconstructed forehand, together with the reference video, and demonstrates the realistic smoothness of the reconstructed 3D motion. The 3D reconstruction of the complete 36 second video is presented in the demonstration video together with the 2D tracking under-pinning the reconstruction.

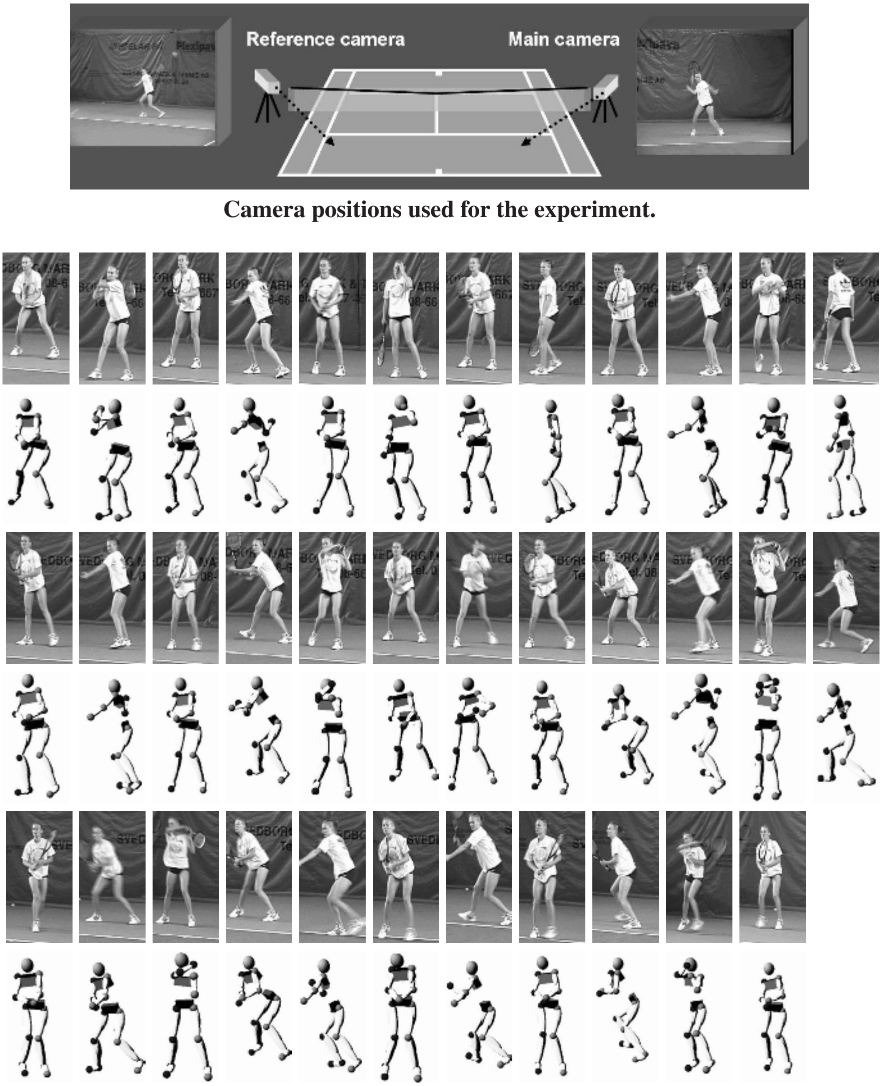
The video and figures 11 and 12 show the qualitative accuracy of our results, and demonstrate that our system can deal with a diverse range of actions recurring over a long sequence. Figure 13 further demonstrates how our system is able to deal with self-occlusion, rapid motion, clutter from the tennis racket, and motion blur.

The system detects outliers as discontinuities in the 3D motion and fills in the missing data via interpolation to form a plausible trajectory. This enables the system to deal with isolated tracking failures. Further, the underlying recognition-based approach to the 2D tracking means the target is freshly detected each frame, and thus ideally placed to recover from 'tracking loss'. In the worst case, the 3D reconstruction will revert to the smooth interpolation from the keyframes (figure 14). How accurate these key poses are depends on how well the sequence is represented by the keyframes, this is specified by the user who defines α the percentage of the sequence which is well-represented by the keyframes.

Our method is well-suited to action sequences with repeated events (e.g. sport). Furthermore, it is possible to quantify the suitability of a sequence for this form of reconstruction by checking how many keyframes are required to represent the desired percentage of the sequence.

8 Closing Remarks

We have presented a method for the 3D reconstruction of articulated body motion from a long monocular sequence. The performance of our system was demonstrated over 36



Camera positions used for the experiment.

Fig. 11. Results of the reconstruction of the entire sequence. Every 50th frame of the 36s long sequence is shown together with the image from our reference camera.

seconds of tennis footage and shown to provide a qualitatively accurate reconstruction. To our knowledge this is longest full-body 3D reconstruction attempted from markerless monocular image data.

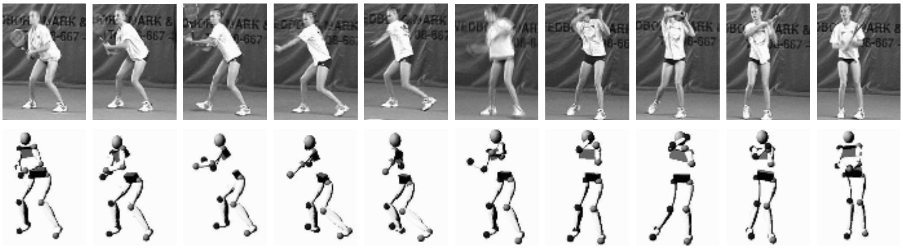
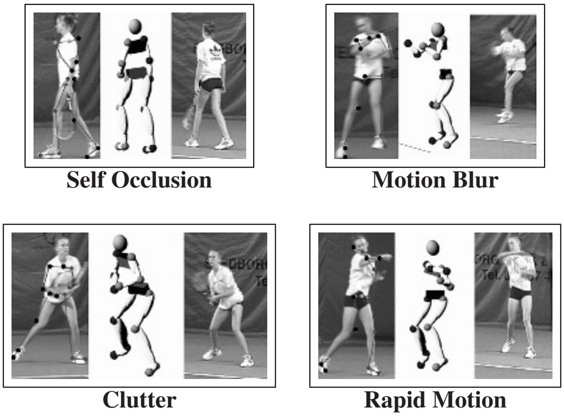


Fig. 12. The reconstruction of one forehand stroke shown together with the images from our reference camera. Note the smoothness of the reconstruction.



Self Occlusion

Motion Blur

Clutter

Rapid Motion

Fig. 13. Examples of reconstructions achieved under difficult imaging conditions. Each case shows the tracked 2D data, the 3D reconstruction from the perspective of the reference camera, and the view from the reference camera.

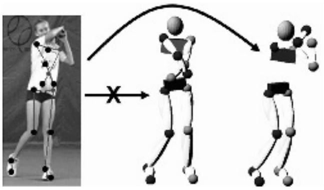


Fig. 14. The importance of maintaining a smooth motion. A large error is encountered in the joint localisation (a). Without enforcing motion smoothness, the frame would be reconstructed as (b). In (c) the reconstructed frame is shown after enforcing smoothness constraints.

References

1. A. Blake and M. Isard. *Active Contours*. Springer, 1998.
2. C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *CVPR*, 1998.
3. J. Deutscher, A. Blake, and I. Reid. Motion capture by annealed particle filtering. *Proc. Conf. Computer Vision and Pattern Recognition*, 2000.
4. V. Lepetit, A. Shahrokni, and P. Fua. Robust data association for anline applications. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2003.
5. G. Loy and A. Zelinsky. Fast radial symmetry for detecting points of interest. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(8):959–973, 2003.
6. G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *Poc of European Conference on Computer Vision*, 2002.
7. T. Moselund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3), 2001.
8. D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2003.
9. K. Shoemake. Animating rotation with quaternion curves. In *SIGGRAPH*, 1985.
10. H. Sidenbladh and M. Black. Implicit probabilistic models of human motion for synthesis and human tracking. In *Poc of European Conference on Computer Vision*, 2002.
11. H. Sidenbladh, M. Black, and D.J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *Poc of European Conference on Computer Vision*, pages 702–718, 2000.
12. C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3d body tracking. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2001.
13. C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2003.
14. J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *Poc of European Conference on Computer Vision*, 2002.
15. C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single image. *Computer Vision and Image Understanding*, 80(3):349–363, 2000.
16. K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *ICCV*, July 2001.
17. L. Zelnik-Manor and M. Irani. Event-based video analysis. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2001.