Synchronizing Image Sequences of Non-Rigid Objects

Phil Tresadern and Ian Reid
Department of Engineering Science, University of Oxford
Parks Road, OX1 3PJ, UK
{pat,ian}@robots.ox.ac.uk

Abstract

For stereopsis, images of a given scene must be captured at the same instant to ensure temporal consistency. For sequences of images (i.e. video streams) this requires the potentially costly and technically complex process of synchronizing cameras. We present a simple but effective method for automatically recovering the sub-frame temporal offset between image sequences taken using *unsynchronized* cameras. Having recovered the offset, we obtain the affine structure of a non-rigid motion. The technique is demonstrated for the application of human motion capture.

1 Introduction

In stereopsis applications, two views of a dynamic, non-rigid scene must be taken at the same instant in order that triangulation be used for scene reconstruction. Corresponding pairs in an image sequence from each viewpoint can then be used to reconstruct the scene at each instant in time. A prime application of this is in human motion capture where typical commercial systems [7] use multiple cameras, synchronized by hardware. While this is neither difficult nor expensive for two or three analogue cameras, there are more complex issues to be dealt with when using many cameras or with current digital technology. Moreover such synchronization cannot be relied upon in many interesting applications involving less controlled environments such as surveillance or sporting analysis.

In this paper we present a simple but effective method of recovering the temporal offset, at sub-frame accuracy, between two sequences of non-rigid motion recorded with unsynchronized cameras (see Fig. 1). We apply the algorithm to sequences of human motion capture data to recover the affine structure of a running person and compare this with ground-truth data to demonstrate the efficacy of the method.

For reconstruction algorithms in which the relief of the scene is small compared to the depth of the scene, it is sensible (both from the point of view of simplicity and in terms of computational accuracy) to approximate the viewing projection as an affinity. In their seminal work Tomasi and Kanade [9] showed that, under affine projection, a $2V \times N$ matrix of N image points in V views (the so-called “measurement matrix”) could be written as a product of a $2V \times 3$ projection matrix and a $3 \times N$ structure matrix. In consequence, they showed that the rank of a noiseless measurement matrix is bounded above by 3 and that the “best” approximation to this could be achieved by factorizing the measurement matrix using the Singular Value Decomposition. Reid and Murray [6] qualified “best”

BMVC 2003 doi:10.5244/C.17.64

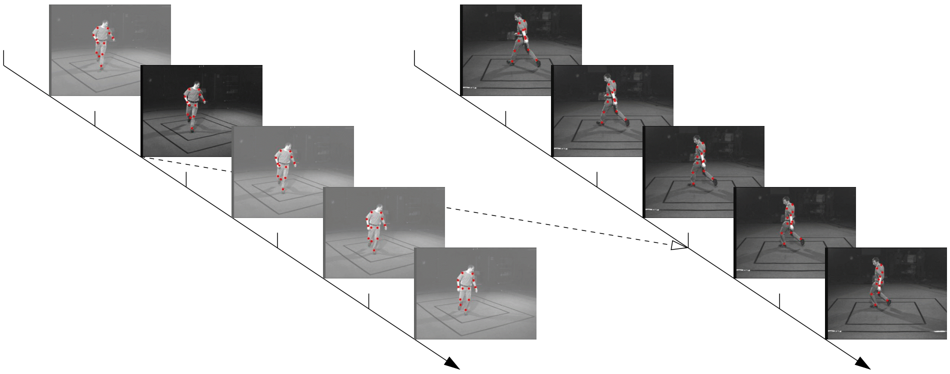

Figure 1: Two sequences of non-rigid motion for which the sub-frame temporal offset is to be recovered, as indicated by the arrow

by showing that the structure and motion computed using the TK factorization method (under suitable assumptions about the errors) minimized the reprojection error and could therefore be interpreted as a maximum likelihood estimate for the structure and motion.

If more than three frames of a rigid body are seen then there exist sufficient constraints to perform self-calibration. This was noted by Tomasi and Kanade (though not in this context) and subsequently formalized by Quan [5] for all affine projection models. In our case we consider only two cameras and non-rigid motion. We therefore have sufficient information to compute affine structure using factorization but in the present work we do not attempt to upgrade this to Euclidean (though see Section 4). The assumption of non-rigid motion is of crucial importance since our algorithm works by identifying the temporal alignment for which the "most consistent structure" can be identified.

Our work is inspired by the similar method proposed by Wolf and Zomet [10]. Since the factorization algorithm of Tomasi and Kanade relies on consistent structure, it cannot be applied to whole sequences of non-rigid motion but can be applied to each pair of frames. Over the sequence of images, the rank 3 constraint is valid only at pairs of corresponding frames where structure is consistent. Wolf and Zomet used this fact to define the "energy" of a measurement matrix above an expected rank bound. This energy is lowest when structure is consistent (i.e. at corresponding frames). In their algorithm, feature correspondence is not assumed between the two sequences although it is assumed that each tracked point in the second sequence can be expressed as a linear combination of points tracked in the first. In the sense that we presently assume the same point set is tracked in each view, our work is less general than theirs but ours is distinguished by the fact that we recover *sub-frame accurate* alignment for sequences of *differing, unknown frame rates*. This temporal offset is then used to recover structure at a higher temporal resolution than either sequence taken individually, with reduced reprojection errors.

The temporal alignment of image sequences has also been extensively studied in [1, 2, 3]. Their methods enable both temporal *and* spatial alignment of sequences, without explicit feature correspondence. Therefore, in some cases sequences can be aligned even without any spatial overlap. However, crucially this work assumes that the images are related by a 2D homography which restricts its application to planar scenes or to scenes viewed by cameras with coincident (or nearly so) optic centres. By contrast, we are concerned with spatially separated cameras viewing 3D scenes.

The remainder of the paper is structured as follows: Section 2 describes the formulation of the problem and outlines the algorithm used for the recovery of the offset. Section 3 discusses implementation details and results while future development of the algorithm and conclusions are presented in Section 4.

2 Problem formulation

2.1 Rank constraints of the measurement matrix

Assuming only two sequences are to be aligned, we can compose a 4×N measurement matrix **W** as follows:

$$\mathbf{W} = \begin{bmatrix} u_1^1 & u_2^1 & \cdots & u_N^1 \\ v_1^1 & v_2^1 & \cdots & v_N^1 \\ u_1^2 & u_2^2 & \cdots & u_N^2 \\ v_1^2 & v_2^2 & \cdots & v_N^2 \end{bmatrix} \tag{1}$$

Where $(u_n^i, v_n^i)^T$ is the vector of image coordinates of the nth feature in the ith view.

This measurement matrix is then normalized with respect to its centroid such that each row has zero mean. As shown by Tomasi and Kanade, assuming affine projection and exact feature correspondence, this new measurement matrix can then be factorized into a 4×3 projection matrix **P** and a 3×N structure matrix **X** such that:

$$\mathbf{W} = \mathbf{PX} = \begin{bmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \ldots & \mathbf{X}_N \end{bmatrix} \tag{2}$$

Where each $\mathbf{P}_i$ is the 2×3 affine projection matrix associated with the ith view and each $\mathbf{X}_n$ is the 3×1 vector of coordinates in 3D space of the nth feature.

It is clear that **W** is rank 3 under ideal conditions. However, under practical conditions the effects of noise and inexact feature correspondence result in **W** having rank >3 (i.e. the fourth singular value is non-zero). When **W** is composed of points measured at the same instant in time, the fourth singular value depends only on the noise in the measurements. However, this rank constraint is violated if **W** is composed from two images corresponding to different instants in time since the structure, **X**, is inconsistent between the two frames and the fourth singular value will be relatively high. The relative magnitude of the fourth singular value can therefore be used as a measure of the synchronization between frames, reaching a minimum at the correct offset. This suggests a straightforward way of obtaining the synchronization as discussed in the following sections.

2.2 Recovery of frame correspondences

Given two sequences of the same dynamic scene, we assume that corresponding frames in the target and reference sequences (f_{tgt} and f_{ref}, respectively) are related by:

$$f_{tgt} = \alpha f_{ref} + \delta t \tag{3}$$

where α is the target:reference frame rate ratio and δt is the offset between the 0th frames in each sequence. We seek to recover δt and in some cases α. We define a *frame window* as a sequence of M consecutive frames from a given sequence. For every frame F in the reference sequence the algorithm attempts to recover the frame, F', corresponding to the same instant in time in a target sequence of T frames as follows:

1. Define the $2\times$MN measurement matrix, **W**, as:

$$\mathbf{W}(f) = \begin{bmatrix} u^F_{ref,1} & \cdots & u^F_{ref,N} & u^{F+1}_{ref,1} & \cdots & u^{F+M-1}_{ref,N} \\ v^F_{ref,1} & \cdots & v^F_{ref,N} & v^{F+1}_{ref,1} & \cdots & v^{F+M-1}_{ref,N} \\ u^f_{tgt,1} & \cdots & u^f_{tgt,N} & u^{f+1}_{tgt,1} & \cdots & u^{f+M-1}_{tgt,N} \\ v^f_{tgt,1} & \cdots & v^f_{tgt,N} & v^{f+1}_{tgt,1} & \cdots & v^{f+M-1}_{tgt,N} \end{bmatrix} \tag{4}$$

   where $(u^k_{ref,n}, v^k_{ref,n})^T$ is the image location of the nth feature in the kth frame of the reference sequence and $(u^k_{tgt,n}, v^k_{tgt,n})^T$ is the image location of the nth feature in the kth frame of the target sequence.

2. Normalize $\mathbf{W}(f)$ with respect to the centroid such that each row has zero mean.

3. Normalize the 4-vector of singular values of $\mathbf{W}(f)$ and define $\mathrm{s}_F(f)$ as the value of the fourth singular value after normalization.

In the case where α is unknown (or known to be non-unity) we should set M = 1 since M $>$ 1 will span different time intervals in each sequence, although for $\alpha \approx 1$, it is reasonable and beneficial to use a small frame window. The aim of the algorithm is then to minimize $\mathrm{s}_F(f)$ over all f in the range 0...T-M. This minimum should occur at frame F', the frame in the target sequence corresponding to frame F in the reference sequence.

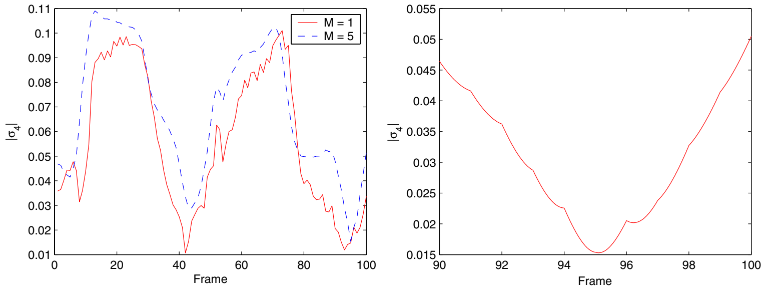

Figure 2: (left) $\mathrm{s}_{95}(f)$ using frame window sizes of 1 frame and 5 frames (right) $\mathrm{s}_{95}(f)$ using interpolated feature locations within the interval [90,100]

Fig. 2(left) shows $\mathrm{s}_{95}(f)$ evaluated using two sequences, part of which can be seen in Fig. 1. A minimum can clearly be seen at frame 95, as desired, although a local minimum is also apparent at frame 42 as a result of the periodic motion of the subject. It can be seen that increasing M reduces the effect of this local minimum. The occurence of local minima in the case of periodic motion was also noted by Wolf and Zomet [10].

2.3 Recovery of synchronization to sub-frame accuracy

Once frame correspondence has been recovered between the sequences robust estimation methods (e.g. RANSAC) can be used to fit a line to the correspondences, recovering α and

δt (the gradient and y-intercept, respectively). However, correspondences between frames are accurate only to within one frame and data is also discarded in the outlier rejection process. Therefore, it is desired to improve these estimates using every frame in the reference sequence at sub-frame accuracy in order to reduce reprojection (and therefore reconstruction) errors.

If the sequences are captured at a sufficiently high frame rate, it is reasonable to assume that the motion of a point from one frame to the next is approximately linear. As a result, f_{tgt} can be estimated for any f_{ref} using Equation 3 and $\mathbf{W}(f)$ can be generated for non-integer values of f by using linear interpolation of the frames either side of f. Evaluation of $s_F(f)$ is then possible for a range of values around the expected value, f_{tgt}, at an arbitrary resolution as shown in Fig. 2(right). As a result, the minimum of $s_F(f)$ can be determined to sub-frame accuracy for every frame in the reference sequence. Improved estimates of α and δt can then be obtained using a least squares linear approximation.

3 Implementation & Results

To compare results with ground truth data, two 300 frame sequences featuring a subject running within a room were captured using accurately calibrated cameras, hardware synchronized at 60Hz. The locations of 12 joints (shoulder, elbow, wrist, hip, knee and ankle on each side) were labelled by hand in each frame of the first and second sequences (see Fig. 1). Except where stated, α was fixed at 1 since the frame rates of the cameras were known to be equal. To establish ground truth the calibration of the cameras was used to "correct" image locations to be consistent with the known epipolar geometry. At no other time was the calibration of the cameras exploited.

3.1 Performance using perfectly matched sequences

Using a pair of perfectly matched sequences (i.e. $\alpha = 1$ and $\delta t = 0$), the algorithm described in Section 2.2 was applied to each reference frame to recover the corresponding frame, F', in the target sequence. Fig. 3 shows a plot of the error, F - F', for the first 100 frames of the 300 frame sequence using 3 frame window sizes. Although in most instances this error is very small, in some cases (e.g. frames 42 and 95) the absolute offset error is large due to the ambiguity caused by local minima (Fig. 2). It can be seen that increasing the size of the frame window reduces errors, as would be expected since the amount of data supplied to the algorithm increases.

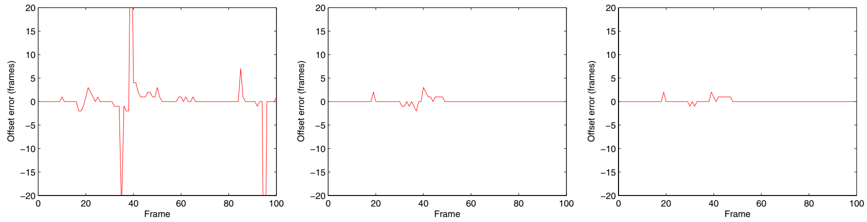

Figure 3: Single frame offset error using frame windows of (left) 1 frame (centre) 3 frames (right) 5 frames

3.2 Performance over varying temporal offset

To test the algorithm at different offsets, unsynchronized sequences were formed by taking interleaved frames from the available synchronized sequences (e.g. sequences with an offset of 0.5 frames were generated by taking frames 0, 2, 4,... from the target sequence and frames 1, 3, 5,... from the reference sequence). For the purposes of this test, the target sequence was created using frames 0, 10, 20,... (an effective frame rate of 6Hz). For some integer k, the reference sequence can be composed of frames k, 10+k, 20+k,... to simulate an offset of $k/10$ frames. The algorithm was applied for the range $k = 0 \dots 10$ and Fig. 4 compares the recovered offsets with the ideal values.

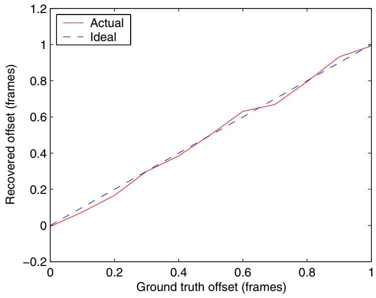

Figure 4: Recovered offset over a range of values

It can be seen that the recovered offsets are typically accurate to within a few hundreths of a frame, a result which is especially encouraging in light of the fact that (i) the assumption of linear motion between frames degrades as frame rate decreases (ii) using every 10th frame drastically reduces the number of points for line fitting. Both of these problems are associated with short sequences of low frame rate whereas a typical application would utilize long sequences at a relatively high frame rate.

3.3 Robustness in the presence of noise

We now show how noise affects the performance of the algorithm when used to recover the offset at sub-frame accuracy. The image points in both the reference and target sequence of a typical, perfectly matched pair were perturbed by random noise distributed as $N(0,\sigma)$ with σ in the range $1 \dots 8$ pixels. The algorithm was then used to determine δt for each σ. Typical results of the initial pass of the algorithm (to determine frame correspondence) are shown in Fig. 5.

It is clear that the number of incorrect frame correspondences increases with noise, up to a point where the desired line fit is no longer detectable. This results in a greater likelihood of an incorrect recovery of δt. Of course, this somewhat naive implementation discards all local minima of $\mathrm{s}_F(f)$ in favour of the global minimum. A better method (currently not implemented) would use not the number of inliers (as in the “standard” RANSAC) but would seek a minimum of $\sum_i s_i(\alpha i + \delta \mathrm{t})$ as the best consensus.

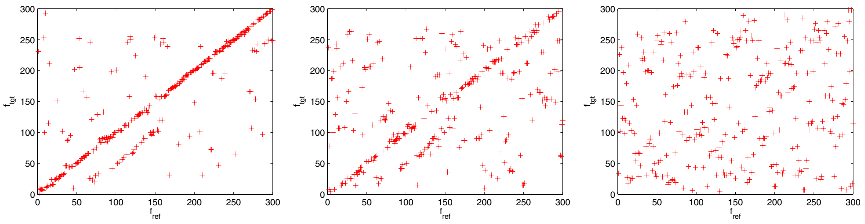

Figure 5: Frame correspondence results with added noise: (left) σ = 2 pixels (centre) σ = 4 pixels (right) σ = 8 pixels

3.4 Reprojection errors

In this work we are considering affine, rather than Euclidean, structure so a direct comparison of the reconstruction errors is not straightforward. Instead, we measure the total reprojection error for the case of (i) interpolation using the known δt and known α, (ii) interpolation using the recovered δt and known α, (iii) interpolation using the recovered δt and recovered α and (iv) no interpolation. We then compare these errors to the reprojection error as a result of assuming affine projection, for exactly matched frames.

In the case where no interpolation is used the measurement matrix is composed of the reference frame and the nearest corresponding frame from the target sequence. This is effectively the case when δt is recovered to the nearest frame, although in the case where the sequences are offset by exactly 0.5 frames this represents the worst case error. The total reprojection error can be trivially computed as:

$$E = ||\mathbf{W}_{groundtruth} - \mathbf{W}_{reprojected}||_F \tag{5}$$

Table 1 shows the four reprojection errors for a typical sequence pair, showing that using sub-frame accuracy offers a substantial improvement in results over single frame accuracy. Fig. 6 shows the affine reconstruction seen from a novel view.

Reprojection method	Total reprojection error
Exactly matched frames	5.1188
Interpolated (known δt, known α)	10.0045
Interpolated (recovered δt, known α)	10.0114
Interpolated (recovered δt, recovered α)	10.4089
No interpolation	14.4676

Table 1: Reprojection errors

3.5 Sequences of differing frame rates

In previous experiments, α was fixed at 1, effectively constraining the minimization when determining δt. We now use sequences of different frame rates and seek to recover δt *and* α best satisfying Equation 3. A target sequence was constructed, consisting of frames 0, 2, 4,... from the first sequence and a reference sequence of frames 1, 4, 7,... from

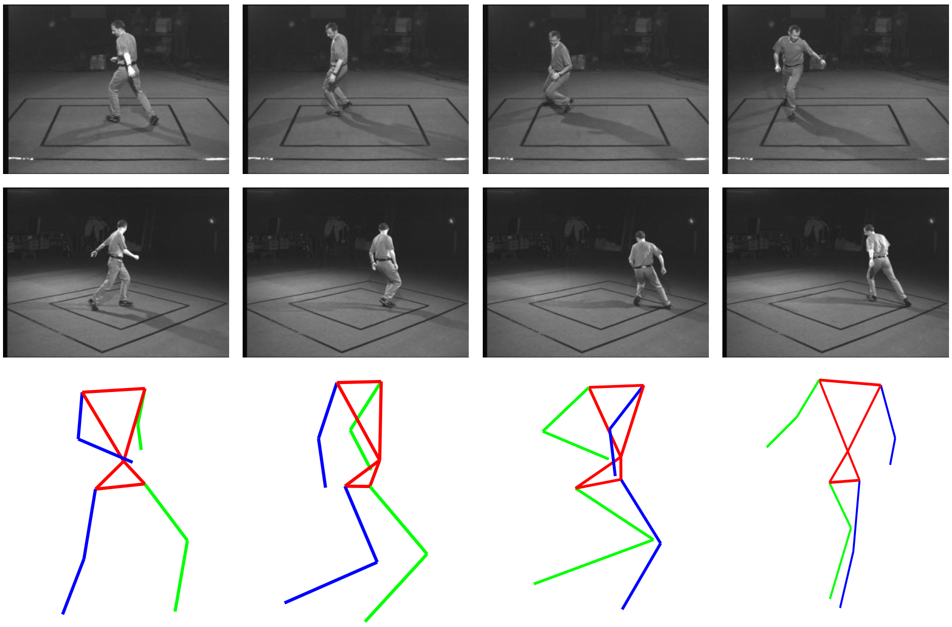

Figure 6: (top) Frames from target sequence (middle) Frames from reference sequence (bottom) Novel view of the affine reconstruction using the recovered offset

the second (i.e. a target:reference frame rate ratio of 1.5 with an offset of 0.5 frames). Applying the algorithm recovered values of 1.4987 and 0.5836 for α and δt, respectively. For comparison, the algorithm was also applied with α fixed at 1.5, recovering an offset of 0.5298.

3.6 Aligning NTSC and PAL

While in the previous experiment we simulated non-unity in α by selectively interleaving frames, here we demonstrate the method for the temporal alignment of a juggling sequence captured from different viewpoints using uncalibrated NTSC and PAL cameras. The reference sequence, captured using a digital camera, consisted of 150 colour frames at 30Hz with a resolution of 320×240 pixels whilst the target sequence, captured using an analogue camera, contained 250 greyscale frames at 25Hz with a resolution of 720×576 pixels, shown in Fig. 7(top). Feature correspondences were determined manually and were not exact since the calibration (and hence epipolar geometry) was unknown. By inspection, it was seen that frame 1 of the reference sequence corresponded approximately to frame 115 of the target sequence. Fig. 8 shows the plot of the minima of $s_F(f)$ for all F from which we recover values of α and δt as 0.8378 ($\approx$ 25/30) and 113.3209, respectively. The parallel bands, offset from the main band, are a result of the local minima due to the periodicity of the juggling motion (see Fig. 2). These bands are less evident from approximately frame 75 onwards, at which point the juggling motion was varied.

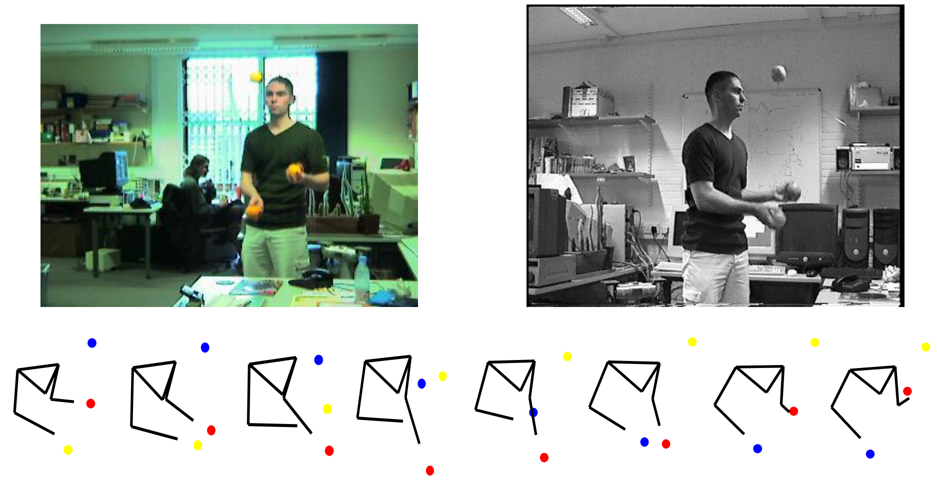

Figure 7: (top) Corresponding frames from the new sequence (bottom) Reconstruction of the juggling sequence seen from a novel view

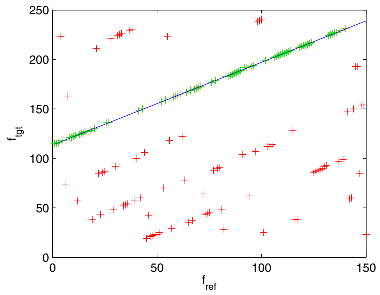

Figure 8: Frame correspondences and the least squares linear approximation

4 Conclusion

We have presented a method for recovering the temporal offset between unsynchronized sequences of non-rigid motion to sub-frame accuracy. Our work improves on [10] in that (i) we determine the offset to sub-frame accuracy and (ii) the algorithm can be applied to sequences of different frame rates. We assess the benefits of determining sub-frame offsets by considering reprojection errors and show that these are significantly reduced. The method is successfully applied to sequences of human motion capture using cameras of different formats (NTSC and PAL).

The algorithm could be extended in a number of ways. For example, the offset could be recovered for more than two sequences by synchronizing each pair separately, under the constraint that the recovered offsets are consistent.

As mentioned in section 1, affine structure is recovered rather than Euclidean (Fig. 6). If the structure of the subject is known, constraints can be placed upon the 3D reconstruction to maintain consistency across the sequence and enable an upgrade to Euclidean structure. For example, it is straightforward to apply the constraint that rigid links are constant in length between frames, as proposed by Liebowitz [4]. Taking this further, a more model-based approach may be applied such that actual ratios of lengths between links are enforced, as in Taylor [8].

5 Acknowledgments

This project is funded by the EPSRC. We are also grateful to Ben Tordoff and Andrew Davison for their helpful support and comments.

References

[1] Y. Caspi and M. Irani. A step towards sequence-to-sequence alignment. In *Proc. 19th IEEE Conf. on Computer Vision and Pattern Recognition, Hilton Head Island*, volume 2, pages 682–689, June 2000.

[2] Y. Caspi and M. Irani. Alignment of non-overlapping sequences. In *Proc. 8th Int'l Conf. on Computer Vision, Vancouver*, volume 2, pages 76–83, July 2001.

[3] Y. Caspi E. Shechtman and M. Irani. Increasing space-time resolution in video. In *Proc. 7th European Conf. on Computer Vision, Copenhagen*, volume 1, pages 753–768, May 2002.

[4] D. Liebowitz and S. Carlsson. Uncalibrated motion capture exploiting articulated structure constraints. In *Proc. 8th Int'l Conf. on Computer Vision, Vancouver*, volume 2, pages 230–237, July 2001.

[5] L. Quan. Self-calibration of an affine camera from multiple views. Technical Report R.T. Imag-Lifia 26, LIFIA - CNRS - INRIA, 1994.

[6] I. D. Reid and D. W. Murray. Active tracking of foveated feature clusters using affine structure. *International Journal of Computer Vision*, 18(1):41–60, April 1996.

[7] Vicon Motion Systems. Online specifications. `http://www.vicon.com/`.

[8] C. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *Proc. 19th IEEE Conf. on Computer Vision and Pattern Recognition, Hilton Head Island*, volume 1, pages 677–684, June 2000.

[9] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[10] L. Wolf and A. Zomet. Correspondence-free synchronization and reconstruction in a non-rigid scene. In *Proc. Workshop on Vision and Modelling of Dynamic Scenes, Copenhagen*, May 2002.