Camera Calibration From Human Motion

Philip A. Tresadern, Ian D. Reid

Abstract

This paper presents a method for the self-calibration of non-rigid affine structure to a Euclidean co-ordinate frame from only two views by enforcing constraints derived from the known structure of the human body, such as piecewise rigidity and approximate symmetry. We show that the proposed algorithm is considerably more efficient yet equally accurate when compared to previous methods. The resulting structure and motion is then refined further using a full bundle adjustment to give maximum likelihood values for body segment lengths and joint angles. A quantitative analysis is presented using synthetic data whilst qualitative results are demonstrated for real examples of human motion.

Key words: Self-Calibration; Human Motion Analysis.

1. Introduction

In order to recover dynamic, non-rigid human motion, commercial systems (*e.g.* [1]) employ a number of hardware-synchronized and accurately calibrated cameras under controlled studio conditions. High contrast markers at anatomical locations on the surface of the body are tracked in each camera, their 3D co-ordinates computed by triangulation and a 'skeleton' fitted to the resulting marker set using kinematic constraints. Various motion parameter (*e.g.* joint angles) can then be computed over the sequence.

A more practical system would eliminate many of these constraints such that human motion can be recovered from stock footage using only a few (*e.g.* two) cameras that are unsynchronized and uncalibrated. This would not only reduce the cost and technical complexity of the solution but could also be employed in applications such as surveillance or sporting analysis. For this to be achieved, however, the system must address four key problems: recovering projected anatomical landmarks; establishing spatial correspondence; camera synchronization; camera calibration (the focus of this paper).

Although the recovery of projected anatomical landmarks (*e.g.* joint centres) has been partially addressed via database searching [2–4], regression [5–7] and assembling kinematic structure from independently detected body parts [8–10], we simply label joint locations by hand since tracking is not our goal in this paper. This intuitive labelling of the image features (*e.g.* "left shoulder") also provides all the information required for matching (*i.e.* spatial correspondence).

Camera synchronization ensures that image features matched between *sequences* also correspond to the same instant in *time* before triangulation. In commercial systems, this is achieved using hardware although several works have shown that the image data itself can provide sufficient constraints to synchronize the cameras [11–13]. In particular, our previous studies have shown this to be the case for sequences of human motion [14,15].

The problem we address in this work, however, is that of camera calibration from only two views using constraints derived from the known structure of the human body, thus providing 3D structure in a Euclidean co-ordinate frame where meaningful motion parameters (*e.g.* joint angles and body segment lengths) may be recovered. Commercial systems employ an explicit calibration procedure prior to motion capture in order to calibrate several cameras with respect to each other. Specifically, a markered 'wand' of known dimensions is waved around the workspace during motion capture such that aggregating all measurements provides a dense set of features with known geometry in the workspace from which the cameras can be calibrated. Alternative approaches have used a sampling strategy to estimate fundamental matrices between pairs of image sequences based on epipolar tangents to frontier points on the silhouette [16].

In constrast, we develop a method that builds on work by Liebowitz and Carlsson [17] and is more suitable for the practical system discussed. In particular, affine structure of the human body is recovered via factorization [18] from only two views, where there are insufficient constraints on the projection matrices to calibrate the cameras. To recover a unique solution, known properties of the human body (in particular, symmetry and piecewise rigidity) are exploited in order to provide further constraints. In other related work, Taylor [19] showed that applying even stronger constraints on structure (specifically, enforcing fixed *ratios* of segment lengths) was sufficient to recover scene structure (up to some depth ambiguities and with user input) for a single image.

In the original implementation, Liebowitz and Carlsson's algorithm [17] conducted a non-linear minimization over both projection and structural constraints. In this work, we improve the method by strictly enforcing projection constraints, thus reducing the dimensionality of the solution space by 66%. We show that this not only results in a considerable increase in efficiency but also eliminates many ambiguities in the original implementation and provides an intuitive initialization for the optimization. We finish by completing a bundle adjustment over all free parameters to minimize a geometric (rather than algebraic) error and recover the maximum likelihood solution. [1]

2. Structure From Motion by Factorization

In their groundbreaking paper, Tomasi and Kanade [18] showed that using an *affine* camera model (a sensible approximation in many cases), projection is *linear* such that a matrix of feature trajectories from a rigid scene can be written as:

$$\mathbf{W} = \begin{bmatrix} \mathbf{x}_1^1 & \cdots & \mathbf{x}_N^1 \\ \vdots & \ddots & \vdots \\ \mathbf{x}_1^V & \cdots & \mathbf{x}_N^V \end{bmatrix} = \begin{bmatrix} \mathbf{P}_1 \\ \vdots \\ \mathbf{P}_V \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_N \end{bmatrix} = \mathbf{PX} \tag{1}$$

where $\mathbf{x}_n^v$ is the image of the 3D feature $\mathbf{X}_n$ under the projection matrix $\mathbf{P}_V$. Using this result, they showed that $rank(\mathbf{W}) \leq 4$ although normalization of the feature trajectories with respect to translation results in a tighter lower bound of 3. Crucially, they noted that $\mathbf{W}$ can then be factorized into $\mathbf{P}$ and $\mathbf{X}$ using the Singular Value Decomposition (SVD) and retaining only the data associated with the relevant singular values. This affine structure and motion was then "upgraded" to a Euclidean co-ordinate frame by applying $2V$ constraints (zero skew and unit aspect ratio) on the rows of $\mathbf{P}$ [20].

3. Self-Calibration

We now discuss the applicability of this method to two sequences of a *non-rigid* scene where we recover structure and motion by factorization independently at each time instant, i. With some abuse of notation, we redefine $\mathbf{P}_i$ as the $4{\times}3$ normalized (with respect to translation) projection matrix at time i and $\mathbf{X}_i$ as the $3{\times}N$ structure matrix at time i (also normalized with respect to translation). At each instant i, structure is known only up to an unknown affine transformation, $\mathbf{G}_i$:

$$\mathbf{W}_i = \mathbf{P}_i\mathbf{X}_i = \mathbf{P}_i\mathbf{G}_i^{-1}\mathbf{G}_i\mathbf{X}_i \tag{2}$$

where each $\mathbf{G}_i$ is an invertible, homogeneous $3{\times}3$ matrix that can be factorized by QR-decomposition ($\mathbf{G}_i \rightarrow \mathbf{Q}_i\mathbf{B}_i$) into a 3D rotation, $\mathbf{Q}_i$, and an upper-triangular matrix, $\mathbf{B}_i$. Since $\mathbf{Q}_i$ effects a rotation of the Euclidean coordinate frame *after calibration* it can be discarded without loss of generality. Consequently, as each $\mathbf{B}_i$ has six independent, non-zero elements a sequence of F frames has $6F - 1$ degrees of freedom, up to a global scale factor.

We define $\Omega_i = \mathbf{B}_i^T\mathbf{B}_i$ such that $\mathbf{B}_i$ is recovered from Ω_i by Cholesky factorization *if and only if* Ω_i *is positive definite.* Eigen-decomposition of $\Omega_i = \mathbf{V}_i\mathbf{D}_i\mathbf{V}_i^T$ such that $\mathbf{B}_i = \mathbf{D}_i^{1/2}\mathbf{V}_i^T$ explains the action of $\mathbf{B}_i$ geometrically as a rotation into a new coordinate frame, followed by an anisotropic scaling.

[1] Preliminary results of this method were published in [15]

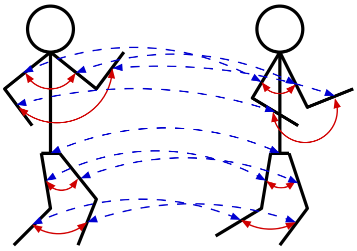


Fig. 1. Symmetry (solid) and rigidity (dashed) constraints between a pair of reconstructions.

3.1. *Motion constraints*

To recover the required set of all $\mathbf{B}_i$ that transforms each affine reconstruction into Euclidean space, constraints are applied to all projection matrices, $\mathbf{P}_i$, in a form of self-calibration [18,20]. Specifically, for a given $\mathbf{B}$ (dropping subscripts for clarity) the axes, $\mathbf{i}^T$ and $\mathbf{j}^T$, of an affine projection matrix transform to $\mathbf{i}^T\mathbf{B}^{-1}$ and $\mathbf{j}^T\mathbf{B}^{-1}$ where the costs for skew, r_{skw}, and difference in length, r_{asp}, are given by:

$$\begin{aligned} r_{skw} &= \mathbf{i}^T\mathbf{B}^{-1}\mathbf{B}^{-T}\mathbf{j} \\ &= \mathbf{i}^T\Omega^{-1}\mathbf{j} && (3) \\ r_{asp} &= \mathbf{i}^T\mathbf{B}^{-1}\mathbf{B}^{-T}\mathbf{i} - \mathbf{j}^T\mathbf{B}^{-1}\mathbf{B}^{-T}\mathbf{j} \\ &= \mathbf{i}^T\Omega^{-1}\mathbf{i} - \mathbf{j}^T\Omega^{-1}\mathbf{j}. && (4) \end{aligned}$$

Under most circumstances, it is sensible to impose constraints that the vectors $\mathbf{i}^T\mathbf{B}^{-1}$ and $\mathbf{j}^T\mathbf{B}^{-1}$ be orthogonal and have unit aspect ratio (*i.e.* $r_{skw} = r_{asp} = 0$). As a result, at a given instant, i, three or more views of the subject provide at least six linear constraints on $\mathbf{B}_i^{-1}\mathbf{B}_i^{-T} = \Omega_i^{-1}$ and a linear least squares solution for Ω_i^{-1} minimizes r_{skw} and r_{asp} [18,20]. However, in this case (where only two views are available) there are insufficient constraints on Ω_i^{-1} and an infinite number of solutions exist.

3.2. *Structural constraints*

In order to overcome this deficiency for fewer than 3 views, it has been shown [17,19] that using knowledge of the human body imposes further constraints on reconconstructions. Figure 1 illustrates the four *symmetry* constraints (solid arrows) between the arms and legs and nine *rigidity* constraints (dashed arrows) on the left/right upper arm, forearm, thigh and foreleg, and hips, as suggested by Liebowitz and Carlsson [17].

More formally, two 3D vectors, $\mathbf{X}_{i,p}$ and $\mathbf{X}_{i,q}$, representing *different* links in the *same* affine reconstruction, i, transform to $\mathbf{B}_i\mathbf{X}_{i,p}$ and $\mathbf{B}_i\mathbf{X}_{i,q}$ in Euclidean space. Likewise, the vectors $\mathbf{X}_{i,p}$ and $\mathbf{X}_{j,p}$ representing the *same* link in *different* affine reconstructions, i and j, constrain both Ω_i and Ω_j. The residual errors, r_{sym} and r_{rig}, are given by:

$$\begin{aligned} r_{sym} &= \mathbf{X}_{i,p}^T\mathbf{B}_i^T\mathbf{B}_i\mathbf{X}_{i,p} - \mathbf{X}_{i,q}^T\mathbf{B}_i^T\mathbf{B}_i\mathbf{X}_{i,q} \\ &= \mathbf{X}_{i,p}^T\Omega_i\mathbf{X}_{i,p} - \mathbf{X}_{i,q}^T\Omega_i\mathbf{X}_{i,q} && (5) \\ r_{rig} &= \mathbf{X}_{i,p}^T\mathbf{B}_i^T\mathbf{B}_i\mathbf{X}_{i,p} - \mathbf{X}_{j,p}^T\mathbf{B}_j^T\mathbf{B}_j\mathbf{X}_{j,p} \\ &= \mathbf{X}_{i,p}^T\Omega_i\mathbf{X}_{i,p} - \mathbf{X}_{j,p}^T\Omega_j\mathbf{X}_{j,p}. && (6) \end{aligned}$$

Although motion and symmetry constraints alone are sufficient for self-calibration at each instant, rigidity constraints (that apply at different instants) account for scale changes over time that are induced by perspective. Further, since rigidity constraints apply between pairs of reconstructions there is a combinatorial number of them, not all independent (*e.g.* $\mathbf{X}_{i,p} = \mathbf{X}_{j,p}$ and $\mathbf{X}_{i,p} = \mathbf{X}_{k,p}$ imply $\mathbf{X}_{j,p} = \mathbf{X}_{k,p}$). Although they may be applied between consecutive instants ($\{0,1\}$, $\{1,2\}$ *etc.*) as in [17], we apply them with respect to the *same* reconstruction ($\{0,1\}$, $\{0,2\}$ *etc.*) in order to prevent the scale from drifting over the sequence.

4. Baseline method

We begin by presenting the 'baseline' method proposed by Liebowitz and Carlsson [17]. It is against this method that we base our comparisons in Section 8.

4.1. *Recovery of local structure*

To recover the rectifying transformations (and hence Euclidean structure and motion), all residuals must be minimized. However, this cannot be achieved using linear methods since motion and structure constrain Ω^{-1} and Ω, respectively. Liebowitz and Carlsson optimize directly over the $6F-1$ elements of all $\mathbf{B}_i$ (up to scale) using a cost function of the form:

$$C = w_{cam} \cdot c_{cam} + c_{str} \qquad (7)$$

where

$$c_{cam} = \sum r_{skw}^2 + \sum r_{asp}^2 \tag{8}$$

$$c_{str} = \sum r_{sym}^2 + \sum r_{rig}^2 \tag{9}$$

and w_{cam} weights the costs according to the relative confidence in the motion and structural constraints. Having recovered all $\mathbf{B}_i$, they compute Euclidean structure and motion at each frame: $\widetilde{\mathbf{X}}_i \leftarrow \mathbf{B}_i\mathbf{X}_i$ and $\widetilde{\mathbf{P}}_i \leftarrow \mathbf{P}_i\mathbf{B}_i^{-1}$, respectively. We refer to this as *local* structure since the choice of coordinate frame is arbitrary at each time instant and rigid transformations between frames are not recovered.

4.2. *Recovery of global structure*

From the enforcement of rigidity over the sequence, any scaling due to perspective over time can be recovered from the computed projection matrices. Therefore, perspective effects over time can be removed by rescaling the image measurements as if viewed orthographically. All F normalized images of N features can then be treated as a *single* image of FN features in a static scene with a common co-ordinate frame. To normalize the data, each Euclidean projection matrix $\widetilde{\mathbf{P}}_i$ is decomposed into its internal and external parameters:

$$\widetilde{\mathbf{P}}_i = \begin{bmatrix} \mathbf{K}_{i,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{i,2} \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{P}}_{i,1} \\ \widehat{\mathbf{P}}_{i,2} \end{bmatrix} \tag{10}$$

where $\widehat{\mathbf{P}}_{i,n}$ is an orthographic projection matrix (such that $\hat{\mathbf{i}}^T\hat{\mathbf{j}} = 0$ and $\hat{\mathbf{i}}^T\hat{\mathbf{i}} = \hat{\mathbf{j}}^T\hat{\mathbf{j}} = 1$) and $\mathbf{K}_{i,n}$ is the corresponding affine calibration matrix of the form:

$$\mathbf{K} = \begin{bmatrix} s & \beta \\ 0 & \kappa s \end{bmatrix} \tag{11}$$

where s is the scale, κ is the aspect ratio and β the skew (subscripts are omitted for clarity). The image measurements are normalized to the same size using the scale factors, s, and a single Ω is recovered for the entire sequence, yielding *global* structure where rotation and relative translation of the body between frames is also recovered. This global structure is then approximated by an articulated body of median segment lengths.

5. Proposed method

Although theoretically sound, the method presented in [17] has a number of practical limitations: it is inefficient since optimization is performed over $6F - 1$ variables; it has no intuitive initialization since linear solutions for Ω_i are seldom positive definite such that the $\mathbf{B}_i$ cannot be recovered by Cholesky decomposition; there is considerable ambiguity when implementing the method since each $\mathbf{B}_i$ can be parameterized in several different ways (our experience suggests this can significantly affect performance); the value of w_{cam} must be chosen empirically.

5.1. *Minimal parameterization*

To address these shortcomings, we propose an improved method that exploits a minimal parameterization of Ω_i based upon reasonable assumptions regarding camera calibration. Specifically, we *strictly* enforce motion constraints, resulting in reconstructions that are constrained to lie in a Euclidean coordinate frame. This has an unambiguous implementation, reduces computational complexity and provides an intuitive starting point for non-linear optimization.

By strictly enforcing motion constraints, we eliminate four degrees of freedom in Ω_i^{-1}. The four motion constraints defined by (3) and (4) yield a linear system with a two dimensional null-space that is spanned by two possible values for Ω_i^{-1} (denoted by $\Omega_{i,1}^{-1}$ and $\Omega_{i,2}^{-1}$). Any linear combination of $\Omega_{i,1}^{-1}$ and $\Omega_{i,2}^{-1}$ satisfies all motion constraints *exactly.* We parameterize all such Ω_i^{-1} using polar coordinates:

$$\begin{aligned} \Omega_i^{-1}(r,\theta) &= r(\cos(\theta)\cdot\Omega_{i,1}^{-1} + \sin(\theta)\cdot\Omega_{i,2}^{-1}) && (12) \\ &= r\cos(\theta)(\Omega_{i,1}^{-1} + \tan(\theta)\cdot\Omega_{i,2}^{-1}) && (13) \end{aligned}$$

such that for any given θ, the eigenvalues of Ω_i^{-1} are equal up to scale for all positive r. Using this parameterization, only $2F - 1$ parameters are required to describe the calibration of the entire sequence (in contrast to the $6F - 1$ non-zero elements of $\mathbf{B}_i$ employed in the baseline method). However, additional measures are required in order to enforce the constraint that Ω_i^{-1} be positive-definite.

5.2. *Optimization*

In an early version of this method, we proposed a simple solution to this problem. Using the polar parameterization of Ω_i^{-1}, we computed the six values of θ for which $|\Omega_i^{-1}| = 0$ (where at least one eigenvalue is zero). The range $[0, 2\pi)$ is therefore divided into six intervals, only one of which corresponds to a positive-definite Ω_i^{-1} for all positive r. This interval, $(\theta_{min}, \theta_{max})$, is recovered by evaluating the eigenvalues of Ω_i^{-1} at the midpoints of the six intervals. The midpoint of $(\theta_{min}, \theta_{max})$ then provides a simple initial value for θ, whilst r is initialized to unity.

Further investigation of the problem (also reported in [21]) revealed that the minimum of r_{sym} can be computed in closed form for every time instant in the sequence to provide an improved initial value of θ. Preliminary investigations suggest there is no closed form solution for the complete system.

We then minimize c_{str} only ($c_{cam} = 0$ by design such that w_{cam} is no longer required) over all $r > 0$ and $\theta \in (\theta_{min}, \theta_{max})$ such that the resulting Ω_i^{-1} are guaranteed to be positive definite and all $\mathbf{B}$ can be recovered by Cholesky factorization. Note that since Ω_i^{-1} is singular at θ_{min} and θ_{max} the cost at these values increases to infinity. As a result, the minimization is effectively 'self-constraining' such that unconstrained methods are successfully employed in all but a few cases.

6. Bundle adjustment

Having recovered local and global structure using the minimal parameterization, we approximate the recovered structure with an articulated model of median segment lengths and estimated pose, as in [17]. We then further optimize all free parameters using a final bundle adjustment (Levenberg-Marquardt, implemented as `lsqnonlin` in Matlab). At this point we relax symmetry constraints since they are the most uncertain of our assumptions.

Minimization of the geometric reprojection error is achieved by optimizing over the v views of i frames for all camera parameters – image scales $\{s_{i,v}\}$, camera rotations $\{\mathbf{R}_v\}$ and translations, $\{\mathbf{t}_v\}$ – and structural parameters – segment lengths, $\mathbf{L}$, and pose parameters, $\{\boldsymbol{\phi}_i\}$. We retain the assumption that the cameras have unit aspect ratio and zero skew.

Defining $\boldsymbol{\epsilon}$ as the vector of reprojection errors over all measurements, we seek to minimize the sum of squared reprojection errors, $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$, over all frames:

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \sum_v \sum_i \sum_n \| s_{i,v}\mathbf{R}_v\mathbf{X}_{i,n}(\mathbf{L}, \boldsymbol{\phi}_i) + \mathbf{t}_v - \mathbf{x}_{i,v,n} \|_F^2 \tag{14}$$

where $\mathbf{X}_{i,n}(\mathbf{L}, \boldsymbol{\phi}_i)$ is the 3D location of the nth feature in the ith frame given the link lengths, $\mathbf{L}$, and pose parameters, $\boldsymbol{\phi}_i$ and $\mathbf{x}_{i,v,n}$ is the corresponding image measurement. This minimization is achieved by iteratively solving:

$$\Delta\mathbf{p} = -(\mathbf{J}^T\mathbf{J} + \lambda\mathbf{I})^{-1}\mathbf{J}^T\boldsymbol{\epsilon}\mathbf{p} \tag{15}$$

for $\Delta\mathbf{p}$ where $\mathbf{p}$ is the vector of all parameters and $\mathbf{J}$ is the Jacobian (matrix of derivatives) of all measurements with respect to the parameters. λ is a regularization parameter to ensure that the step size remains within the trust region where the linearization, upon which Levenberg-Marquardt is based, remains valid. Since scale and pose parameters are frame dependent, $\mathbf{J}$ is sparse and minimization is computationally efficient. The end result is an articulated model of fixed link lengths, fitted to the anthropomorphic dimensions of the subject (up to scale) and capturing the pose at every frame such that *all* constraints are strictly enforced.

7. Outlier rejection

As with all methods based on linear least squares minimization, gross outliers in joint locations have a highly detrimental effect on algorithm performance. We note, however that simple measures can be taken to eliminate many gross outliers using random sampling methods [22] to estimate the (affine or projective) fundamental matrix. In the case of affine projection, it has been shown that computationally cheap subspace-based methods can be employed to verify spatial matching [23].

We take a different approach based on full perspective projection: the cameras in our application are fixed with respect to each other such that all image pairs in an entire sequence must share the same epipolar geometry. Although at each time instant it is possible to use an affine approximation (since a person's relief is typically much smaller than the viewing distance), motion towards and away from a camera induces perspective effects over the sequence that we can use to our advantage. Each putative feature match in an entire sequence constrains the epipolar geometry and we use this large feature set to estimate the fundamental matrix robustly using RanSaC. The benefits resulting from this procedure are demonstrated in Section 8.1.5.

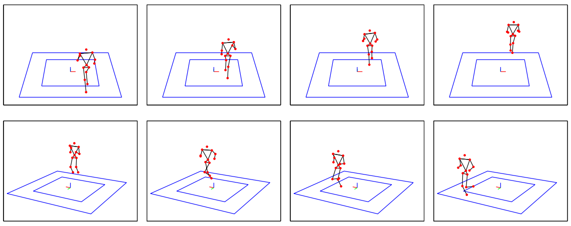
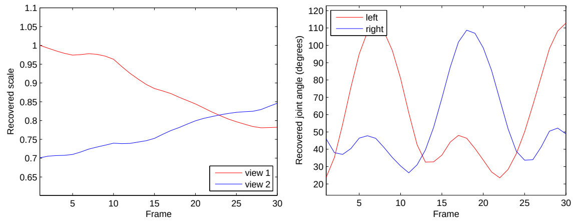


Fig. 2. Synthetic 'running' sequence as seen from two wide baseline viewpoints. The red circles indicate point features used as inputs to the synchronization algorithm.



(a) (b)

Fig. 3. (a) Recovered scaling as a result of perspective effects. (b) Recovered trajectories of the knees during running sequence. The expected periodicity and phase difference is clearly evident.

8. Performance evaluation

We now present results using synthetic data to demonstrate the benefits of the proposed method over the baseline implementation [17].

8.1. *Running sequence*

Two views of a short running motion (consisting of 30 frames) were synthesized using motion capture data from a commercial system (Figure 2). An articulated model of known segment lengths was imaged under perspective projection and the projected image features used to recover affine structure by factorization. Metric structure and motion was then recovered using four methods: (i) rectification using a local implementation of Liebowitz and Carlsson's method ('L&C'); minimal parameter rectification with (ii) no bundle adjustment ('Minimal'); (iii) affine bundle adjustment ('A.B.A.'); (iv) perspective bundle adjustment ('P.B.A.', a 'gold standard' for comparison). This particular sequence was selected since the translation of the subject induced scaling over time due to perspective.

Figure 3a shows the recovered scales as a results of perspective – the subject runs toward one camera and away from the other. The recovered angles at the knees are shown in Figure 3b where the periodicity and phase difference of the running motion is clearly observable.

Table 1
Performance comparison of four methods where it is clear that the minimal parameterization heavily outperforms the original parameterization. Bundle adjustment reduces the errors further at some computational cost.

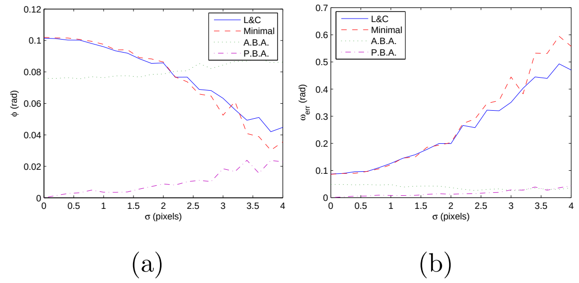
		L&C	Minimal	A.B.A.	P.B.A.
A	# iterations	15	6	6	6
	Time (sec)	1.14	0.20	0.20	0.20
B	# iterations	439	5	5	5
	Time (sec)	2.31	0.039	0.039	0.039
C	# iterations	-	-	9	185
	Time (sec)	-	-	8.934	133.5
Total time (sec)		4.33	1.60	12.36	137.3
Reproj. error (pixels)		1.41	1.44	0.785	$<10^{-3}$
Joint angle error (rad)		0.0521	0.0511	0.0328	$<10^{-3}$
Limb length error (%)		0.958	0.996	0.798	$<10^{-3}$

8.1.1. *Comparison of algorithm efficiency*

Table 1 compares the described methods using noiseless data, based upon (i) number of iterations required for convergence, (ii) time taken (using a 3.2GHz Pentium 4 desktop computer) for convergence, (iii) total time taken (including fixed overhead costs) and (iv) final RMS reprojection error. We show separate metrics for the recovery of local structure (A), recovery of global structure (B) and bundle adjustment (C).

Minimal parameterization clearly outperforms the baseline method in efficiency with little penalty in accuracy while bundle adjustment increases accuracy further at some additional computational cost (although the tradeoff between accuracy and computational expense is controlled via the stopping criterion). As expected, perspective bundle adjustment converges to an almost exact solution with noiseless data. Since structure is recovered up to a rotation and scaling, measuring distances between a recovered joint location and its corresponding ground truth value in 3D is non-trivial. Instead, we use invariant metric quantities such as joint angles and normalized limb length in order to evaluate the quality of the recovered solution.

In the following experiments, we added zero-mean Gaussian noise of increasing standard deviation σ in order to quantify sensitivity. Each algorithm was



(a) (b)

Fig. 4. (a) Recovered ψ (rad) with respect to image feature locations corrupted by noise of standard deviation σ pixels; (b) Recovered ω_{err} (rad) with respect to image feature locations corrupted by noise of standard deviation σ pixels.

applied in order to recover a number of system parameters and errors quantified for each algorithm. This was repeated 20 times for each value of σ in order to estimate the error distribution (although only the means of the distributions are presented for the sake of clarity).

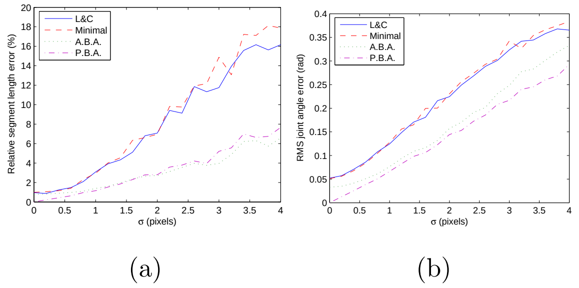
8.1.2. *Recovery of camera parameters*

To compare the recovered rotation between cameras, we recover external parameters from the computed projection matrices. Using the axis-angle notation, a rotation matrix is represented by the unit axis of rotation, $\mathbf{a}$, and angle of rotation, ω, about this axis. We denote ground truth values by $\mathbf{a}_{gt}$ and ω_{gt}, respectively, quantifying error using the angle between axes $\mathbf{a}$ and $\mathbf{a}_{gt}$, $\psi = \cos^{-1}(\mathbf{a}_{gt}^T \mathbf{a})$, and the difference in angle of rotation, $\omega_{err} = |\omega_{gt} - \omega|$.

Figure 4 shows that bundle adjustment results in a considerable reduction in ω_{err} when compared to the 'raw' output of the 'Lieb' and 'Minimal' algorithms. Furthermore, it can be seen that the error following bundle adjustment is relatively invariant to the level of noise. More interestingly, we see that while ω_{err} increases with noise, ϕ decreases (albeit by a smaller amount) for the non-bundle adjusted methods. However, we note that there is an inherent difficulty in quantifying the error between two rotation matrices – an intuitive single error value does not exist and using two values may result in the error compensation that is evident for the non-bundle adjusted data shown here.

8.1.3. *Recovery of segment lengths*

To compare segment lengths, we recover metric 3D structure over the entire sequence and compute the median length for each body segment. These median values are then normalized such that the hips have unit length before comparing them with ground truth values. Since our minimal parameterization strictly enforces motion constraints we might expect a deterioration in the recovered structure (which 'absorbs' all of the measurement errors). However, our results suggest that this effect is very slight.



(a) (b)

Fig. 5. (a) Mean percentage error in recovered limb lengths with respect to image feature locations corrupted by noise of standard deviation σ pixels; (b) Mean RMS error in joint angle (rad) over the knee and elbow joints with respect to image feature locations corrupted by noise of standard deviation σ pixels.

Figure 5a shows mean percentage errors in recovered body segment length using the four methods. We see that error increases sharply with image noise since even a small amount of noise may result in a large *percentage* error in projected length for frames where the limb is almost normal to the image plane. We also see that the error is actually greater for perspective bundle adjustment than for affine bundle adjustment, in disagreement with intuition. However, observation of the error variances (not shown) suggests that this difference is unlikely to be significant.

8.1.4. *Recovery of joint trajectories*

We now show how image noise affects RMS error in joint angle, using the elbow and knee joints that are invariant to global coordinate frame. Figure 5b shows error increase sharply since even a small error in projected length is interpreted as a large error in joint angle. The converse problem is encountered in model-based tracking where rotations out of the image plane are almost unobservable since they result in small image motion [24].

8.1.5. *Sensitivity to gross outliers*

Finally, we investigate the sensitivity of the algorithm to gross outliers as a result of tracking error. Such errors have two deleterious effects: (i) increased RMS projection errors and consequent increased errors in recovered structure; (ii) more seriously, they often result in failure of the algorithm to converge to

Table 2
Convergence frequency, RMS reprojection error and limb lengths error with outliers.

	Convergence	Reproj. error	Limb error	
Method		RMS (pixels)	Mean (%)	Max. (%)
No outliers	100%	1.78	2.52	6.31
Known	100%	1.81	2.71	6.55
RanSaC	81%	2.23	4.36	9.56
Naïve	31%	7.30	5.11	12.10

a sensible solution. We show that such problems are significantly reduced using robust matching techniques.

Using a different synthetic sequence of 38 frames, we added Gaussian noise ($\sigma = 2$ pixels) and performed self-calibration (‘No outliers’). We then deliberately corrupted approximately 10% of the correspondences (selected randomly) with Gaussian noise of standard deviation 40 pixels to simulate gross error and performed self-calibration three more times: (i) after removing all known outliers (‘Known’); (ii) after removing outliers detected using robust matching (‘RanSaC’); (iii) after removing none of the outliers (‘Naïve’). Since this experiment concerns only the early stages of the algorithm, no bundle adjustment was used.

Table 2 shows the convergence frequency over 100 tests, and the RMS reprojection and structure errors averaged over the tests that did converge (only points labelled as inliers were used to compute these values). Methods ‘Naïve’ and ‘Known’ respectively show that performance is poor with outliers present but improves dramatically when they are all removed. The ‘RanSaC’ method shows that robust matching methods [22] provide some defence against such outliers. In particular the percentage of trials that converge is dramatically increased, as well as an expected decrease in structural error.

However, one weakness of binocular outlier rejection schemes is that only those outliers lying far from their estimated epipolar line are detected. Large noise components parallel to the epipolar line remain undetected and continue to influence the recovered structure and motion adversely. Further mitigation against these effects could be obtained using, for example, smooth motion priors to detect remaining outliers.

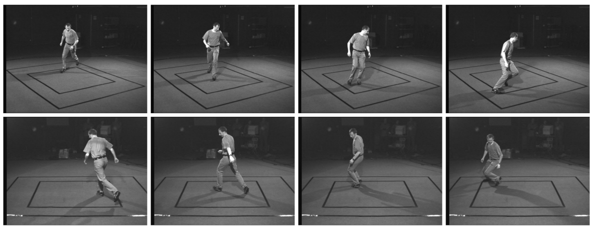
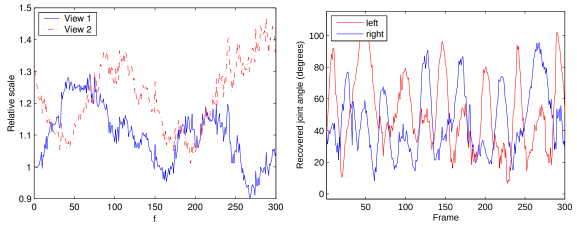


Fig. 6. Running sequence as seen from two wide baseline viewpoints.



(a) (b)

Fig. 7. (a) Recovered scaling as a result of perspective effects. (b) Recovered trajectories of the knees during running sequence. The expected periodicity and phase difference is clearly evident.

Table 3
Recovered body segment lengths (relative to the hips) for the running sequence. The recovered limbs are approximately symmetric and in proportion.

Limb	Left	Right
Upper arm	1.223	1.249
Lower arm	1.004	1.071
Upper leg	1.619	1.679
Lower leg	1.693	1.709

9. Real examples

9.1. *Running sequence*

Applying the algorithm to a real ‘running’ sequence (Figure 6), the affine reconstructions were calibrated using the minimal parameterization in 37 iterations, taking approximately 4.3 seconds. In contrast, Liebowitz’s method took 38 seconds to compute local structure and did not converge on global structure within 10^4 iterations. Affine bundle adjustment was then applied to the recovered structure reducing RMS reprojection error from 5.44 pixels to 2.76 pixels. For comparison, perspective bundle adjustment reduced RMS reprojection error to 2.24 pixels.

Figure 7a shows the recovered scaling of the body

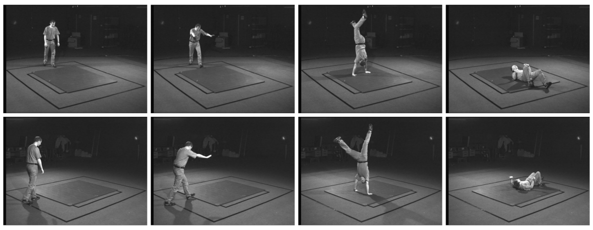
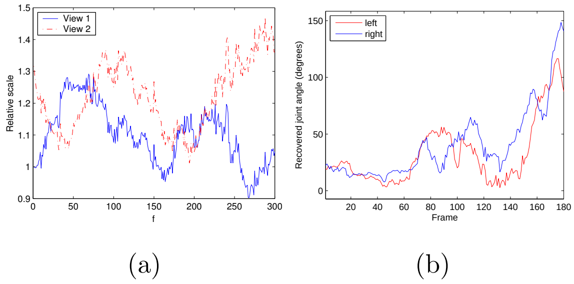


Fig. 8. Handstand sequence as seen from two wide baseline viewpoints.



(a)

(b)

Fig. 9. (a) Recovered scaling as a result of perspective effects. (b) Recovered trajectories of the knees during handstand sequence shoing now periodicity or particular phase difference.

Table 4
Recovered body segment lengths (relative to the hips) for the handstand sequence. The recovered limbs are approximately symmetric and in proportion.

Limb	Left	Right
Upper arm	1.076	1.105
Lower arm	0.856	0.968
Upper leg	1.645	1.719
Lower leg	1.458	1.584

as a result of perspective whilst Figure 7b shows the joint angle trajectories of the knees over 150 frames of the running sequence. The anticipated periodicity and phase difference in the running motion is clearly evident. Table 3 shows the recovered body segment lengths (again, normalized such that the hips have unit length). It can be seen that the recovered body model is in proportion and approximately symmetric, despite the fact we impose no constraints on the symmetry of the body during bundle adjustment.

9.2. *Handstand sequence*

For the 'handstand' sequence (Figure 8), our method converged in 109 iterations, taking only 9.5 seconds, with an RMS reprojection error of 6.79 pixels. Affine bundle adjustment reduced RMS reprojection error to 3.92 pixels, compared with 3.41 pixels following perspective bundle adjustment. In contrast, Liebowitz's method required 6951 iterations, taking 101 seconds, with an RMS reprojection error of 7.56 pixels.

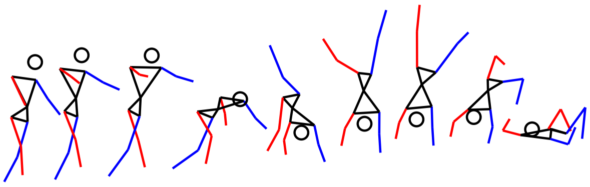


Fig. 10. Euclidean reconstruction of a handstand sequence

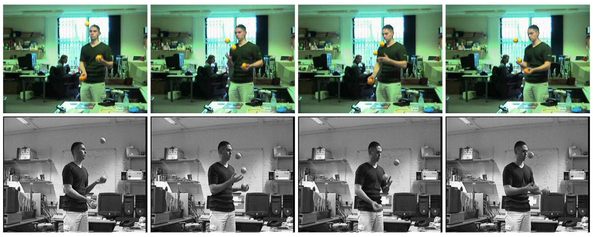
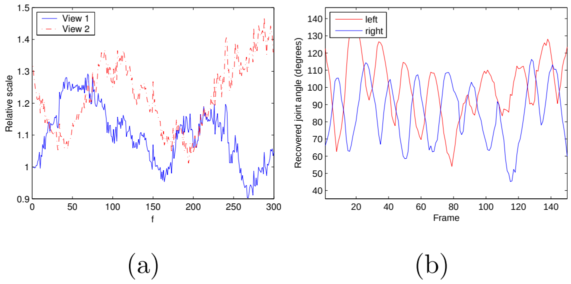


Fig. 11. Juggling sequence as seen from two wide baseline viewpoints.



(a)

(b)

Fig. 12. (a) Recovered scales where we see little change since the subject was not moving with respect to the camera. (b) Recovered trajectories of the elbows during juggling where the out of phase periodic motion is clearly observable.

Figure 9a shows the recovered scales due to perspective and Figure 9b shows the joint angle trajectories of the knees. In this case, there is no periodicity or phase change since the motion is not cyclic. Again, we see that the recovered kinematic structure (Table 4) is in proportion and approximately symmetric. The resulting Euclidean reconstruction of the handstand motion is shown in Figure 10.

9.3. *Juggling sequence*

For the juggling sequence (Figure 11), the minimal parameterization converged in 19 iterations, taking approximately 0.8 seconds, with an RMS

Table 5
Recovered limb lengths (relative to the left upper arm) for the juggling sequence. The recovered limbs are approximately symmetric and in proportion.

Limb	Left	Right
Upper arm	1.000	1.032
Lower arm	0.984	0.982

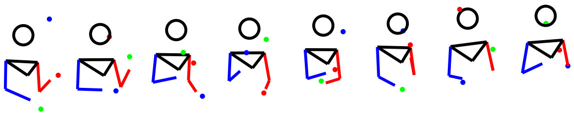


Fig. 13. Euclidean reconstructions from juggling sequence

reprojection error of 4.13 pixels. In contrast, Liebowitz's method required 1425 iterations, taking 20.2 seconds, albeit with a better RMS reprojection error of 3.78 pixels. Affine bundle adjustment reduced RMS reprojection error further to 2.13 pixels, compared with 2.15 pixels following perspective bundle adjustment.

Again, Figure 12a shows the scales due to perspective effect that are small in this case since the subject does not move towards or away from the camera. This lack of change in depth would explain why perspective bundle adjustment performed no better than the affine bundle adjustment for this sequence. Figure 12b shows the recovered joint trajectories of the elbows during the motion where the periodicity of the motion is clearly apparent in addition to the phase difference. Table 5 shows the recovered body segment lengths where we see that the symmetry has been recovered and the segments are in proportion, despite the reduced number of structural constraints (the lengths are normalized with respect to the upper left arm). Figure 13 shows the reconstructed upper body in a Euclidean co-ordinate frame.

10. Conclusion

We have presented a self-calibration method for recovering non-rigid structure and motion in a Euclidean co-ordinate frame from a pair of uncalibrated cameras. The method is an improvement on the original algorithm of Liebowitz and Carlsson [17] in that our alternative parameterization of the solution space is computationally efficient, has an unambiguous implementation and provides an intuitive initialization for the optimization process. A full bundle adjustment over the free parameters then recovers the maximum likelihood solution. Affine bundle adjustment with appropriate scaling resulted in performance almost matching perspective bundle adjustment with large savings in computation. The method is demonstrated on a number of sequences of human motion (both synthetic and real) where the accurate recovery of underlying structure and joint angles is observed.

The key limitation of the method (as shared by [17]) is the need for spatial correspondence. However, we note that a number of algorithms exist for the automatic recovery of joint centre projections that may be applied for this task. Furthermore, the calibration method is strictly a batch process (since it uses all affine reconstructions simultaneously) and could not be employed for real-time applications. An obvious extension would be to develop a recursive process that converges to the maximum likelihood solution. Finally, the sharp increase in joint angle error with noise suggests that integration with a motion model would also be beneficial.

Acknowledgment

This project was funded by the Engineering and Physical Sciences Research Council (EPSRC). Image sequences were provided by Oxford Metrics. Data used in these experiments was obtained from mocap.cs.cmu.edu and created with funding from NSF EIA-0196217. The authors would also like to thank the reviewers for their helpful comments.

References

[1] Vicon Motion Capture Solutions, Online specifications, http://www.vicon.com.

[2] J. Sullivan, S. Carlsson, Recognizing and tracking human action, in: Proc. 7th European Conf. on Computer Vision, Copenhagen, 28–31 May, Vol. 1, Springer LNCS 2350, 2002, pp. 629–644.

[3] B. Stenger, A. Thayananthan, P. H. S. Torr, R. Cipolla, Filtering using a tree-based estimator, in: Proc. 9th Int'l Conf. on Computer Vision, Nice, 14–17 October, Vol. 2, 2003, pp. 1063–1070.

[4] G. Shakhnarovich, P. Viola, T. Darrell, Fast pose estimation with parameter sensitive hashing, in: Proc. 9th Int'l Conf. on Computer Vision, Nice, 14–17 October, Vol. 2, 2003, pp. 750–759.

[5] R. Rosales, S. Sclaroff, Inferring body pose without tracking body parts, in: Proc. 19th IEEE Conf. on Computer Vision and Pattern Recognition, Hilton Head Island, SC, USA, 13–15 June, Vol. 2, 2000, pp. 721–727.

[6] A. Agarwal, B. Triggs, Recovering 3D human pose from monocular images, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (1) (2006) 1–15.

[7] C. Sminchisescu, A. Kanaujia, Z. Li, D. Metaxas, Discriminative density propagation for 3D human motion estimation, in: Proc. 23nd IEEE Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June, Vol. 1, 2005, pp. 390–397.

[8] P. Felzenszwalb, D. Huttenlocher, Pictorial structures for object recognition, International Journal of Computer Vision 61 (1).

[9] D. Ramanan, D. A. Forsyth, A. Zisserman, Tracking people by learning their appearance, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1).

[10] R. Ronfard, C. Schmid, B. Triggs, Learning to parse pictures of people, in: Proc. 7th European Conf. on Computer Vision, Copenhagen, 28–31 May, Vol. 4, Springer LNCS 2353, 2002, pp. 700–714.

[11] I. Reid, A. Zisserman, Goal-directed video metrology, in: Proc. 4th European Conf. on Computer Vision, Cambridge, 15–18 April, Vol. 2, Springer LNCS 1065, 1996, pp. 647–658.

[12] Y. Caspi, D. Simakov, M. Irani, Feature-based sequence-to-sequence matching, International Journal of Computer Vision 68 (1).

[13] T. Tuytelaars, L. V. Gool, Synchronizing video sequences, in: Proc. 22nd IEEE Conf. on Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July, 2004, pp. 762–768.

[14] P. Tresadern, I. Reid, Synchronizing image sequences of non-rigid objects, in: Proc. 14th British Machine Vision Conf., Norwich, 9–11 September, Vol. 2, 2003, pp. 629–638.

[15] P. Tresadern, I. Reid, Uncalibrated and unsynchronized human motion capture : A stereo factorization approach, in: Proc. 22nd IEEE Conf. on Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July, Vol. 1, 2004, pp. 128–134.

[16] S. N. Sinha, M. Pollefeys, L. McMillan, Camera network calibration from dynamic silhouettes, in: Proc. 22nd IEEE Conf. on Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July, Vol. 1, 2004, pp. 195–202.

[17] D. Liebowitz, S. Carlsson, Uncalibrated motion capture exploiting articulated structure constraints, International Journal of Computer Vision 51 (3) (2003) 171–187.

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

[19] C. J. Taylor, Reconstruction of articulated objects from point correspondences in a single uncalibrated image, Computer Vision and Image Understanding 80 (3) (2000) 349–363.

[20] L. Quan, Self-calibration of an affine camera from multiple views, International Journal of Computer Vision 19 (1) (1996) 93–110.

[21] R. Wang, W. K. Leow, Human posture sequence estimation using two un-calibrated cameras, in: Proc. 16th British Machine Vision Conf., Oxford, 5–8 September, 2005.

[22] P. H. S. Torr, D. W. Murray, The development and comparison of robust methods for estimating the fundamental matrix, International Journal of Computer Vision 24 (3) (1997) 271–300.

[23] L. Zelnik-Manor, M. Irani, Degeneracies, dependencies and their implications in multi-body and multi-sequence factorization, in: Proc. 21st IEEE Conf. on Computer Vision and Pattern Recognition, Madison, WI, USA, 16–22 June, Vol. 2, 2003, pp. 287–293.

[24] D. Morris, J. Rehg, Singularity analysis for articulated object tracking, in: Proc. 17th IEEE Conf. on Computer Vision and Pattern Recognition, Santa Barbara, CA, USA, 23–25 June, 1998, pp. 289–297.