Uncalibrated Motion Capture Exploiting Articulated Structure Constraints

David Liebowitz and Stefan Carlsson
Computational Vision and Active Perception Laboratory
Royal institute of Technology, Sweden

Abstract

We present an algorithm for 3D reconstruction of dynamic articulated structures, such as humans, from uncalibrated multiple views. The reconstruction exploits constraints associated with a dynamic articulated structure, specifically the conservation over time of length between rotational joints. These constraints admit metric reconstruction from at least two different images in each of two uncalibrated parallel projection cameras. The algorithm is based on a stratified approach, starting with affine reconstruction from factorization, followed by rectification to metric structure using the articulated structure constraints. The exploitation of these specific constraints allows reconstruction and self-calibration with fewer feature points and views compared to standard self-calibration. The method is extended to pairs of cameras that are zooming, where calibration of the cameras allows compensation for the changing scale factor in a scaled orthographic camera. Results are presented in the form of stick figures and animated 3D reconstructions using pairs of sequences from broadcast television. The technique shows promise as a means of creating 3D animations of dynamic activities such as sports events.

1. Introduction

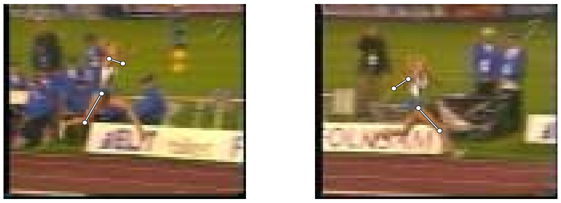
Systems for motion capture , i.e. dynamic 3D reconstruction of articulated structures such as humans in motion, typically employ a large set of cameras in a laboratory environment in order to maximise stability of reconstruction and handle the problem of self-occlusion. These systems are pre-calibrated, so the issue of motion capture is essentially data collection and reconstruction by triangulation. In real world situations such as sports events, on the other hand, the number of cameras viewing a specific situation is often limited to around two or three. The necessity of obtaining both an overview of the situation and close up details further limits the number of cameras useful for motion capture. These cameras are rarely pre-calibrated, which in itself would present problems since they are often hand-held and moving around. In general these cameras are placed at very different positions in order to get complete coverage. This is beneficial for the problem of multiview reconstruction since it implies large baselines between cameras. On the other hand this aggravates the problem of occlusion since cameras are essentially looking at different parts of the human body in motion. Considering the fact that events such as team sports naturally imply occlusions due to multiple actors, the limited number of visible features in multiple views can be a serious problem. A system for motion capture from real world sports events therefore has to face the problems of self-calibration and reconstruction from a limited number of views together with limitations of the number of simultaneously visible feature points in multiple views.

Camera self-calibration from multiple views of general unconstrained structures requires at least 3 views for perspective cameras [11]. Sporting events are most often monitored by zooming cameras from relatively long distances, which implies that the camera geometry is very close to parallel projection. This further aggravates the problem of calibration, increasing the minimum number of cameras from 3 to 4 [15]. It is therefore critical to exploit all available scene constraints for calibration and reconstruction.

Human bodies have traditionally been modelled as "rigid link" articulated structures. Knees, shoulders and elbows are rotational joints connected with (generally) rigid links. The lengths of body segments between rotational joints therefore remain constant over time (see figure 1). We call this the fixed length or rigid link constraint, and it is the purpose of this paper to demonstrate that this constraint can be exploited in an efficient way in order to obtain metric reconstruction and camera calibration from two affine views of human motion.

The specific constraints associated with articulated structures and motion capture has received some attention in earlier work. Much of the research in motion capture has focused on monocular sequences. Leventon and Freeman[8], for example, take a statistical approach by using a motion capture database to build a probability model for motion sequences. Pose of the body in a novel sequence is then inferred by finding the most probable 3D configuration given the imaged co-ordinates. Taylor[19] assumes known size body segments imaged by a scaled orthographic camera, and computes 3D orientation of the segments from foreshortening. Webb [21] considers the rotation axis of relative joints to be fixed for short times and is able to recover 3D structure from monocular sequences of articulated motion.

The paper proceeds with a description of the affine rectification transformation that defines a metric reconstruction of a scene. This is followed in section 3 by a general formulation of the relative length constraint on rectification, including degenerate configurations, and describes how it is used in motion capture. Finally, section 4 presents the rigid link constraint applied to motion capture from broadcast video for both fixed and zooming cameras.



(a) (b)

Figure 1: Body segments, such as the upper arm between the shoulder and elbow, and the upper leg, between the hip and knee, have a constant length when a person moves.

2. From affine reconstruction to metric using scene constraints

The general stratified approach to scene reconstruction from multiple perspective cameras distinguishes projective, affine and metric reconstructions[2]. It is assumed here that an affine reconstruction is known, and must be rectified to metric. The affine reconstruction may be obtained directly from affine cameras. Alternatively, a projective reconstruction from *perspective* cameras. can be upgraded to affine by various methods[14, 4, 10]. This section describes the decomposition of the metric rectifying affine transformation into similarity and 'pure' affine components, and the simplification that results.

Formally, a world point $\mathbf{W}$ and point $\mathbf{A}$ in an affine reconstruction are related by an affine transformation of 3D:

$$\mathbf{W} = \mathsf{H}\mathbf{A} + \mathbf{t} \qquad (1)$$

where H is a full rank 3×3 matrix and $\mathbf{t}$ is a translation 3-vector.

In reconstruction for graphics purposes, similarity transformations may often be neglected. The overall scale, rotation and translation between the world and reconstruction can be chosen arbitrarily. The significant transformation is that which rectifies the reconstruction to within a similarity transformation of the world – a *metric* reconstruction. To this end, the affine rectifying transformation, which has 12 degrees of freedom, is decomposed into two transformations. First, we neglect the translation vector, and then apply RQ decomposition[5] to H_A. H may thus be written as a product of an orthonormal (rotation) matrix S and an upper triangular matrix U:

$$\mathsf{H} = \mathsf{S}\mathsf{U} \qquad (2)$$

In addition, U may be regarded as a homogeneous matrix in the sense that its overall scale factor may be ignored. This scale factor is the isotropic scaling of the similarity transformation $s\mathsf{S}$. U is thus a five degree of freedom 'purely' affine transformation. Determining U and applying it to the affine structure results in metric structure. Any metric constraints on the world coordinates will therefore automatically induce constraints on the rectification U. To see this consider the rectified vector:

$$\mathbf{W}_i = \mathsf{S}\mathsf{U}\mathbf{A}_i + \mathbf{t} \qquad (3)$$

Metric properties of a set of points can always be expressed in terms of inner products:

$$(\mathbf{W}_i-\mathbf{W}_j)^\top(\mathbf{W}_k-\mathbf{W}_l) = (\mathbf{A}_i-\mathbf{A}_j)^\top\mathsf{U}^\top\mathsf{U}(\mathbf{A}_k-\mathbf{A}_l) \quad (4)$$

A similarity structure invariant can likewise always be expressed as a ratio of inner products. Any such ratio implies a linear constraint on the matrix $\mathsf{U}^\top\mathsf{U}$

$$(\mathbf{W}_i - \mathbf{W}_j)^\top(\mathbf{W}_k - \mathbf{W}_l)(\mathbf{A}_i - \mathbf{A}_j)^\top\mathsf{U}^\top\mathsf{U}(\mathbf{A}_k - \mathbf{A}_l) -$$
$$(\mathbf{W}'_i - \mathbf{W}'_j)^\top(\mathbf{W}'_k - \mathbf{W}'_l)(\mathbf{A}'_i - \mathbf{A}'_j)^\top\mathsf{U}^\top\mathsf{U}(\mathbf{A}'_k - \mathbf{A}'_l) = 0$$

We will write

$$\Omega = \mathsf{U}^\top\mathsf{U} \qquad (5)$$

for this matrix, which is homogeneous and has five degrees of freedom. This is closely related to the absolute conic of projective geometry which has the property that it is invariant to 3D similarity transformations. It thus encodes only the affine distortion of the reconstruction, not the similarity component [3]. From Ω the rectification U can easily be computed by Cholesky decomposition[7].

Typical metric constraints involve lengths and angles for point configurations. In the following sections we will exploit constraints such as conservation of relative length over time that are valid for dynamic articulated structures. Internal parameters and auto-calibration constraints may also be included in the computation of the rectification Ω.

3. Rectification constraints from relative length

This section describes a constraint on the metric rectification of an affine reconstruction derived from a known ratio of world 3D lengths. It will be shown that the metric length of a line segment in an affine reconstruction can be expressed in terms of Ω, and that this leads to a linear constraint on the elements of Ω from two segments. Degenerate conditions are described and the application of the constraint to motion capture is introduced.

3.1. The relative length constraint

Consider two 3D line segments in an affine reconstruction, segment $\mathcal{L}_{\mathbf{A}}$ with endpoints $\mathbf{A}_1$ and $\mathbf{A}_2$, and segment $\mathcal{L}_{\mathbf{B}}$ with endpoints $\mathbf{B}_1$ and $\mathbf{B}_2$. The metric lengths of $\mathcal{L}_{\mathbf{A}}$ and $\mathcal{L}_{\mathbf{B}}$, $l_{\mathbf{A}}$ and $l_{\mathbf{B}}$, may be expressed by transforming the endpoints with (3) and taking inner products

$$l_{\mathbf{A}}^2 = (\mathsf{U}\mathbf{X}_{\mathbf{A}})^{\mathsf{T}}(\mathsf{U}\mathbf{X}_{\mathbf{A}}) = \mathbf{X}_{\mathbf{A}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{A}} \qquad (6)$$

where $\mathbf{X}_{\mathbf{A}} = \mathbf{A}_1 - \mathbf{A}_2$.

If the metric length ratio of $\mathcal{L}_{\mathbf{A}}$ and $\mathcal{L}_{\mathbf{B}}$ is λ, $l_{\mathbf{A}}^2 = \lambda^2 l_{\mathbf{B}}^2$ and from (6)

$$\mathbf{X}_{\mathbf{A}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{A}} = \lambda^2\mathbf{X}_{\mathbf{B}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{B}} \qquad (7)$$

where Ω is the symmetric 3×3 matrix with elements

$$\Omega = \begin{pmatrix} \Omega_1 & \Omega_2 & \Omega_4 \\ \Omega_2 & \Omega_3 & \Omega_5 \\ \Omega_4 & \Omega_5 & \Omega_6 \end{pmatrix} \qquad (8)$$

Expanding (7) in terms of the elements of Ω,

$$\begin{aligned} \mathbf{X}_{\mathbf{A}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{A}} - \lambda^2\mathbf{X}_{\mathbf{B}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{B}} = 0 & \\ (X_{\mathbf{A}1}^2 - \lambda^2 X_{\mathbf{B}1}^2)\Omega_1 + 2(X_{\mathbf{A}1}X_{\mathbf{A}2} - \lambda^2 X_{\mathbf{B}1}X_{\mathbf{B}2})\Omega_2 & \\ +(X_{\mathbf{A}2}^2 - \lambda^2 X_{\mathbf{B}2}^2)\Omega_3 + 2(X_{\mathbf{A}1}X_{\mathbf{A}3} - \lambda^2 X_{\mathbf{B}1}X_{\mathbf{B}3})\Omega_4 & \\ +2(X_{\mathbf{A}2}X_{\mathbf{A}3} - \lambda^2 X_{\mathbf{B}2}X_{\mathbf{B}3})\Omega_5 + (X_{\mathbf{A}3}^2 - \lambda^2 X_{\mathbf{B}3}^2)\Omega_6 & = 0 \end{aligned}$$

This is a linear equation in the elements of Ω, which we may write in vector form as

$$\mathbf{c}^{\mathsf{T}}\boldsymbol{\Omega}_{\mathsf{v}} = 0 \qquad (9)$$

where $\mathbf{c} = (c_1, \ldots, c_6)^{\mathsf{T}}$ is the vector of coefficients and $\boldsymbol{\Omega}_{\mathsf{v}} = (\Omega_1, \ldots, \Omega_6)^{\mathsf{T}}$ is the vector of elements of Ω.

Five independent constraints are required to solve for Ω. Given five or more independent constraints, the coefficient vectors $\mathbf{c}_1$ to $\mathbf{c}_n$ can be combined in a constraint matrix. It follows from (9) that

$$\left(\mathbf{c}_1 \ldots \mathbf{c}_n\right)^{\mathsf{T}}\boldsymbol{\Omega}_{\mathsf{v}} = \mathsf{C}\boldsymbol{\Omega}_{\mathsf{v}} = \mathbf{0} \qquad (10)$$

and $\boldsymbol{\Omega}_{\mathsf{v}}$ is the null vector of C, which is of rank five.

In the presence of noise and with more than five constraints, C will generally be of rank six. In this case, singular value decomposition (SVD) provides an estimate of the null vector. The singular vector associated with the smallest singular value is the estimate of the null vector of C that minimises $||\mathsf{C}\boldsymbol{\Omega}_{\mathsf{v}}||$ subject to $||\boldsymbol{\Omega}_{\mathsf{v}}|| = 1$.

3.2. Degeneracies

Constraint degeneracy occurs when a particular constraint fails to provide information about Ω, or when a set of constraints are linearly dependent in (9). Three such conditions are enumerated below.

Parallel line segments. Length ratio in parallel directions is an affine invariant, so parallel line segments in an affine reconstruction preserve their metric length ratio. It is therefore unsurprising that no constraint on Ω is obtained from the relative length of parallel segments. Explicitly, if $\mathcal{L}_{\mathbf{A}}$ and $\mathcal{L}_{\mathbf{B}}$ are parallel with length ratio λ, $\mathbf{X}_{\mathbf{A}} = \lambda\mathbf{X}_{\mathbf{B}}$ (see figure 2 (a)). and (7) becomes

$$\mathbf{X}_{\mathbf{A}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{A}} - \lambda^2\mathbf{X}_{\mathbf{B}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{B}} = \lambda^2\mathbf{X}_{\mathbf{B}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{B}} - \lambda^2\mathbf{X}_{\mathbf{B}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{B}} = 0$$

and is satisfied for all Ω. The elements of the coefficient vector $\mathbf{c}$ will all be zero.

Pairs of parallel line segments. Two constraints obtained from the relative lengths of segment *pairs* that are parallel are not independent. Suppose we have $\mathcal{L}_{\mathbf{A}}$ and $\mathcal{L}_{\mathbf{B}}$ with metric length ratio λ, and $\mathcal{L}_{\mathbf{C}}$ and $\mathcal{L}_{\mathbf{D}}$ with ratio κ, shown in figure 2 (b), we can write two linear constraints on Ω:

$$\begin{aligned} \mathbf{X}_{\mathbf{A}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{A}} - \lambda^2\mathbf{X}_{\mathbf{B}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{B}} &= 0 \\ \mathbf{X}_{\mathbf{C}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{C}} - \kappa^2\mathbf{X}_{\mathbf{D}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{D}} &= 0 \end{aligned}$$

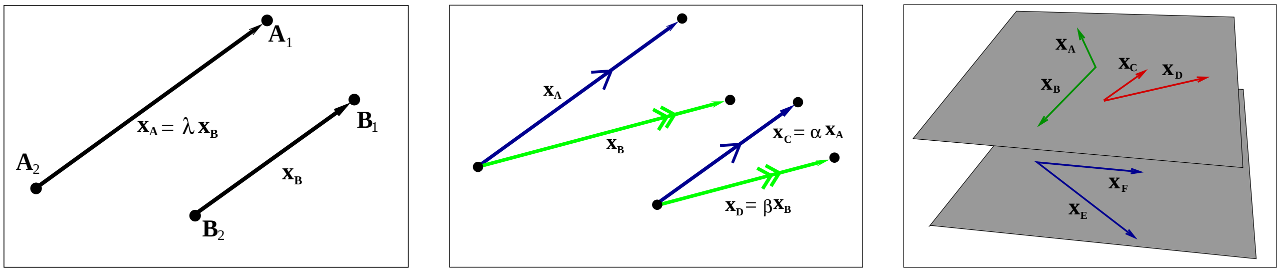
However, if $\mathcal{L}_{\mathbf{A}}$ is parallel to $\mathcal{L}_{\mathbf{C}}$ and $\mathcal{L}_{\mathbf{B}}$ is parallel to $\mathcal{L}_{\mathbf{D}}$, $\mathbf{X}_{\mathbf{C}} = \alpha\mathbf{X}_{\mathbf{A}}$ and $\mathbf{X}_{\mathbf{D}} = \beta\mathbf{X}_{\mathbf{B}}$. Additionally, for parallel directions length ratios are preserved, so $\frac{\alpha}{\beta} = \frac{\kappa}{\lambda}$. The pair of constraints becomes

$$\begin{aligned} \mathbf{X}_{\mathbf{A}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{A}} - \lambda^2\mathbf{X}_{\mathbf{B}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{B}} &= 0 \\ \alpha^2\mathbf{X}_{\mathbf{A}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{A}} - \alpha^2\lambda^2\,\mathbf{X}_{\mathbf{B}}^{\mathsf{T}}\Omega\mathbf{X}_{\mathbf{B}} &= 0 \end{aligned}$$

These two equations are linearly dependent and only one constraint on Ω is obtained.

Co-planar or parallel plane line segments. Any number of constraints originating with coplanar line segments or line segments in parallel planes, as in figure 2 (c), provide only two independent constraints on Ω. Geometrically, the constraints fully define metric structure on a plane [9], and indeed on the pencil of parallel planes, but provide no information about other directions.

As always with degenerate conditions, care must be taken in implementing the constraints. Theoretically degenerate constraints may appear to provide valid constraints due to measurement noise, and near degenerate conditions provide unstable results.



(a) (b) (c)

Figure 2: Degenerate relative length constraints. (a) Parallel line segments provide no constraint on Ω. (b) Parallel pairs of line segments provide one constraint on Ω. (c) Coplanar or parallel plane line segments provided at most two constraints on Ω.

3.3. The rigid link constraint

The relative length constraint can be applied in practice without quantitative knowledge of length ratios. This is the case when the qualitative observation can be made that objects of equal length are present in a scene, determining a unit length ratio constraint. Accordingly, human motion is treated below as the movement of an articulated, rigid link structure. Two cases will be considered:

1. two fixed general affine cameras with different parameters observing an entire motion sequence, and
2. n pairs of scaled orthographic cameras with changing parameters, each imaging one frame of the motion.

In the first case, sequences captured by a pair of unchanging cameras are equivalent to two images of the motion. Each view is regarded as having viewed the entire set of sampled postures from the motion in time in one image. The postures are treated as a static scene, requiring a single reconstruction and rectification transformation. Considered as a static 3D object, the set of postures then contains a number of fixed length objects: each body segment, such as the upper arm, remains a constant length through the set. This is the rigid link constraint.

When the cameras are changing, however, the postures cannot be regarded as a static set in a single pair of images. Each frame must be reconstructed and rectified separately. In this case, rigid links are used for each reconstruction as well as between reconstructions to constrain all the rectifications.

4. Motion capture

The application of the rigid link constraint to motion capture is shown here in the context of broadcast footage of sports. In sports broadcasts, interesting events are typically shown from a second camera in the replay, providing two views of the action. The cameras at sports events are also often placed far from the participants, at the edge of the field or mounted on the stands, and use long focal lengths. The affine camera model is thus applicable, and an affine reconstruction can be created directly from matched points.

This section describe the manual matching process, reconstruction and calibration steps used to rectify this reconstruction using rigid links.

The initial method described assumes a pair of fixed cameras. This is a somewhat naive assumption for sports broadcasts, since cameras often pan and zoom while following a player. Despite this, however, a fixed camera assumption is valid in some situations, and can also result in realistic animation even when violated. This is followed by a method of jointly calibrating a number of pairs of cameras using rigid links. It is shown that, using the scaled orthography camera model, the change in scale factor resulting from a zooming camera can be eliminated, although rotation between cameras remains unknown.

4.1. Affine reconstruction

The first step in affine reconstruction from two views is to select the corresponding image points defining the body parts of interest. A 15 point model of the human skeleton is used, specifying the dominant joints: shoulders, elbows, hips, knees and ankles, as well as the tips of the feet and the head or neck. An example, two views of a triplejump event, appears in figure 3, with the skeleton points shown in two images.

This footage is captured by a pair of cameras some distance from the athlete. The athlete remains approximately the same size and in the same position in the images throughout the sequence. The cameras are thus treated as a pair of fixed affine cameras viewing a motion as if on a treadmill. Translation of the athlete is lost, but the relative movement of the body parts can be reconstructed.

The skeleton points are selected by hand in the images, frame by frame. Two difficulties typically occur. The first is occlusion, either self-occlusion of joints as a body moves, or occlusion from other people or objects in the scene. The second difficulty is motion blur. Rapidly moving features, like the hands, often smear in the images, making precise joint location difficult. Uncertain points in either view are flagged in the clicking process.

Having selected the corresponding point pairs, an initial affine reconstruction is computed using the factorization algorithm of Tomasi and Kanade[20]. Factorization provides a robust and computationally efficient method of computing affine structure from point correspondences using singular value decomposition. It has also been shown to give a Maximum Likelihood reconstruction under Gaussian noise assumptions for image points[16].

Points flagged as uncertain in the selection process can be neglected from this reconstruction. The affine fundamental matrix can then be computed from the cameras and used to recompute the uncertain points by projecting them to the epipolar line defined by the fundamental matrix and the corresponding point in the other view[18]. The system thus tolerates points uncertain in one view, but clearly visible in the other.

An affine reconstruction of the triplejump sequence is depicted in figure 4. Wireframe models for four frames corresponding to figure 3 are shown, with horizontal translation introduced. Note the effect of affine distortion, particularly on the aspect ratio of the body.

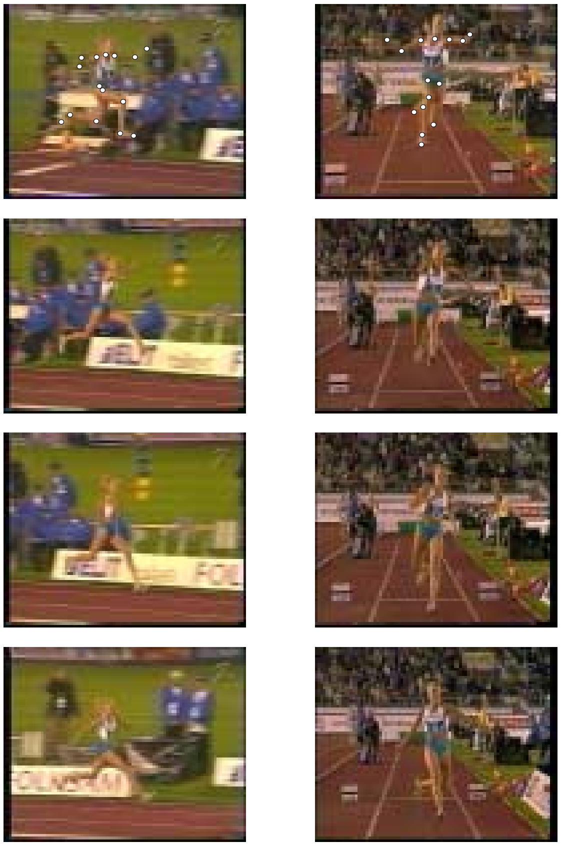


Figure 3: The triplejump. Frames 13, 26, 30 and 51 of an 93 frame sequence, shown in side view above and front view below. Manually selected joints are shown in the first frame of each set.

The affine reconstruction can now be metric rectified using the rigid link constraint.

4.2. Metric Rectification

The rigid link constraint is present in two forms in human motion. This first is the constant length of body segments – the upper and lower arms and legs and the distance between the hip joints – through the motion. There are nine constraints on Ω from each view pair for these particular segments. The second source of constraints is symmetry, the equal length of left and right arm and leg segments, providing four constraints per frame. Constraints are obtained both these sources using (7) with a length ratio of 1, resulting in a constraint matrix of the form in (10). In principal, two frames are sufficient to over constrain Ω, but for interesting motions there are typically tens or hundreds of frames, so Ω is significantly over constrained.

Note also that, in applying (7), the rigid link constraints between frames can be written for the reconstruction of a single body segment between any two postures. There are, for a single segment appearing in n frames, $\binom{n}{2}$ constraints, of which $n-1$ are independent. The approach taken here is to write the $n-1$ constraints for each segment in successive pairs of frames in time.

SVD of the constraint matrix and Cholesky decomposition of the resulting estimate of Ω yield a rectification matrix U, and metric 3D structure follows from (3).

The final step in reconstructing the motion is regularization of segment sizes. Since there is noise in the system, the metric humanoid figure does not have exactly fixed length body segments throughout the reconstructed sequence. The model for each frame of the sequence is thus replaced by a fixed size humanoid with median segment lengths while preserving the angles of body segments. The results appear in figure 5.

Stick figures representations of motion have limited usefullness on their own. A solid skeletal model provides a far more informative representation of the motion. Ultimately, of course, the movement can be mapped to a realistic, textured humanoid model. Accordingly, the motion capture data is exported to a format understood by a commercial animation package, the *Character Studio* plugin to *3D Studio Max*. The generic biped triplejumper appears in figure 6.

The quality of the reconstruction of motion is difficult to evaluate. Clearly, the accuracy of calibrated, marker based, multi-camera systems cannot be replicated. However, when viewed as an animation, the motion appears to be a faithful reproduction of the action. The animation is smooth and duplicates not only the coarse motion of the arms and legs,

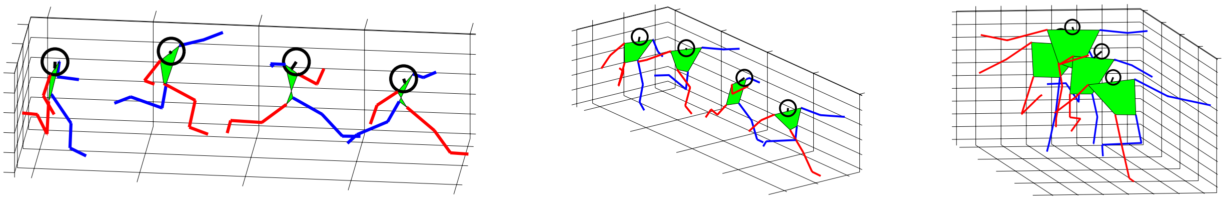


Figure 4: Four postures from the affine reconstruction of the triplejump sequences. The same four postures are shown from three viewpoints. Compare to the corresponding frames in figure 3.

but also captures some of the subtle elements of the movement, such as the asymmetry in the hip and shouder positions through the jump.

4.3. Changing cameras

The cameras filming sports events like football commonly zoom and pan (rotate) to follow a player or the ball. In many situations, the rotation is nececessarily limited to keep the player in the field of view, since the camera is some distance away. There is, however, a significant translation in the image co-ordinates of body points.

The effect of zooming is to increase the imaged size of the player. Consider the example of figure 7, which shows four frames from a sequence of a football goal, with the cameras zooming noticably. It will now be shown how the effect of zooming can be eliminated from the images by calibrating the set of cameras.

Given two sequences acquired with changing cameras, the matched points in the pair of images at each time instance must be be reconstructed separately. The affine reconstructions obtained from *each* frame pair are related to the world by a general affine transformation, including a similarity component. There is thus an arbitrary rotation, translation and scaling in addition to the affine distortion encoded by the rectification Ω relating the *metric* rectifications of each frame. It is, however, still possible to write rigid link constraints on the rectification of each reconstruction. These are constraints on the rectifications for each reconstruction, denoted $\Omega^{(j)}$, $j = 1, \ldots, m$ for a pair of sequences of length m.

In the changing camera case, the rigid link constraint can be applied in two contexts. First, symmetry constraints that the right and left arm and leg segments are the same length are valid in each reconstruction. There are thus at most four constraints on each $\Omega^{(j)}$ of the form (9).

Second, the rigid link constraint may be applied across reconstructions:

$$(\mathbf{X}^j)^\top \Omega^{(j)} (\mathbf{X}^j) - (\mathbf{X}^k)^\top \Omega^{(k)} (\mathbf{X}^k) = 0 \qquad (11)$$

where $\mathbf{X}^j$ and $\mathbf{X}^k$ represent the same rigid body segment in reconstructions j and k. This leads to a linear constraint on the vector of elements of both $\Omega^{(j)}$ and $\Omega^{(k)}$ similar in form to (9).

Up to nine constraints of this form can be written for each pair of reconstructions from the arms, legs and hips. Each $\Omega^{(j)}$ is thus over constrained. Forming a vector of elements of all the $\Omega^{(j)}$'s, the rectification of all reconstructions can be computed from the SVD of a constraint matrix whose rows are obtained from (11).

In practice, since there are far fewer constraints on each rectification than is the case for the single rectification with fixed cameras, it is useful to introduce camera parameter constraints. Assuming that the cameras may be modelled as scaled orthographies, the rows of each projection matrix are orthgonal and of equal length. For a camera $\mathsf{P}^j = (\mathbf{p}_1^\top, \mathbf{p}_2^\top)$, these constraints are[20]

$$\mathbf{p}_1^\top \Omega^{-1} \mathbf{p}_2 = 0 \text{ and } \mathbf{p}_1^\top \Omega^{-1} \mathbf{p}_1 - \mathbf{p}_2^\top \Omega^{-1} \mathbf{p}_2 = 0$$

providing four linear constraints on each $(\Omega^{(j)})^{-1}$. Since these constraints are linear in the inverse of each $\Omega^{(j)}$, they cannot be simply combined with the rigid link constraints. A solution is to use an iterative method to minimize a cost function over all the rectification affinities $\mathsf{U}^{(j)}$. The minimization can be initialized by a linear solution for the set of $\Omega^{(j)}$'s or from the fixed camera method.

Note that the relative scale of metric reconstructions in the set is explicitly constrained by the rigid link constraints. The metric reconstructions are thus up to a common scale, and the set of resulting $\mathsf{U}^{(j)}$'s returns m metric reconstructions with a common scale factor. However, since each initial affine reconstruction is a general affine transformation away from world structure, each has its own unknown Euclidean rectification component. That is, there remains a set of unknown rotations and translations between metric reconstructions of individual frame pairs.

Now, since the scaling is global to all the reconstructions, the relative scaling between pairs of metric cameras is known. The effect of a zooming camera can thus be eliminated by scaling the image points in each sequence relative to the first camera. Assuming negligable rotation of the cameras through the sequence, it is possible to return to the fixed camera method having removed the effect of zooming, and thus compute a metric reconstruction of the *entire* motion. Figure 8 shows skeleton image points in frames 1, 30 and 40 and the image points in frames 30 and 40 scaled according to the recovered camera parameters. Applying the fixed camera algorithm to the scaled image data yields

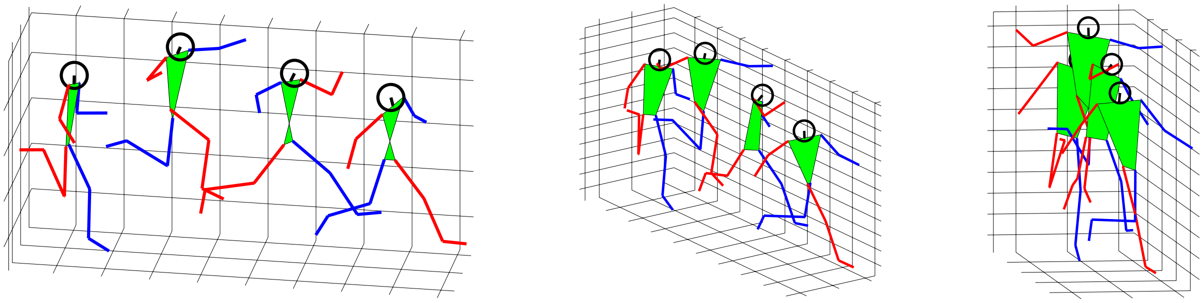


Figure 5: Four postures from the final metric reconstruction of the triplejump sequences. The same four postures are shown from three viewpoints. Compare to the corresponding frames in figure 4.

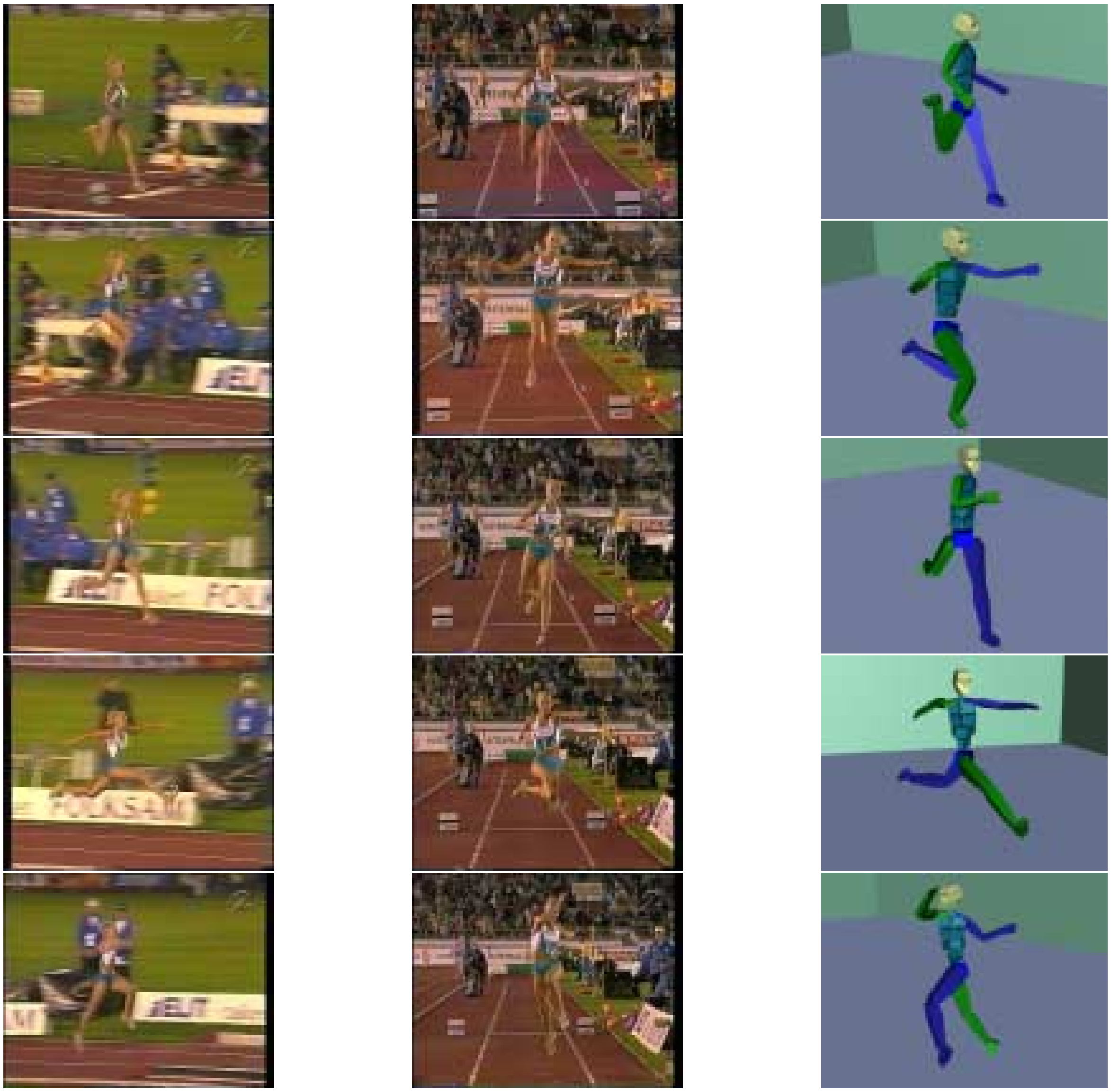


Figure 6: Five frames from the triplejump, equally spaced in time. In each row, frames from the two input image sequences appear alongside snapshots of an animated generic biped from various viewpoints.

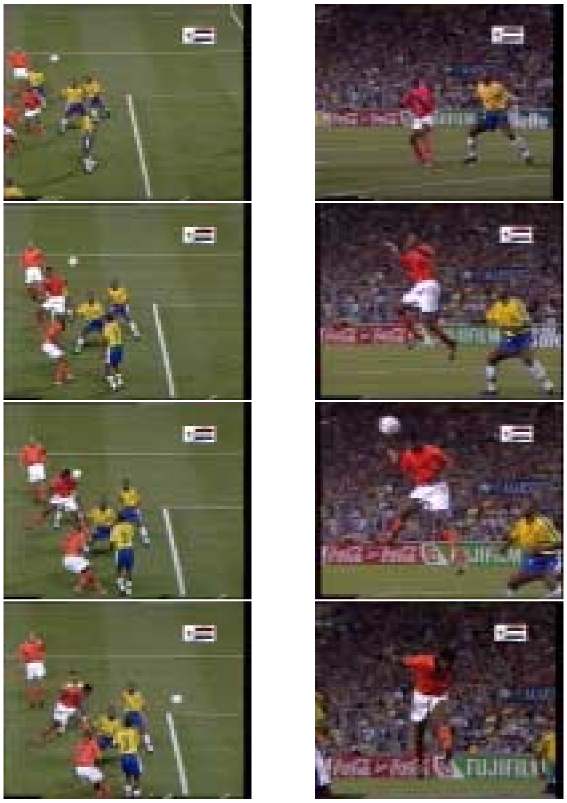


Figure 7: A football goal. Frames 1, 20, 30 and 40 of a 42 frame sequence, shown in two views.

the animated biped shown in 9.

Note that the effects of panning cannot be dealt with in this way, since the orientation between different pairs of cameras is not recovered. This would require some static scene information.

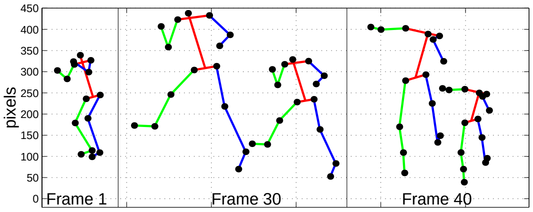


Figure 8: Zoom compensation in images from frames 1, 30 and 40 from the second camera in the football example. The selected image points for frame 1 appear on the left. In the middle and on the right (for frames 30 and 40) are large figures which are plots of the selected image points in those frames, and smaller figures obtained by scaling according to the recovered camera parameters.

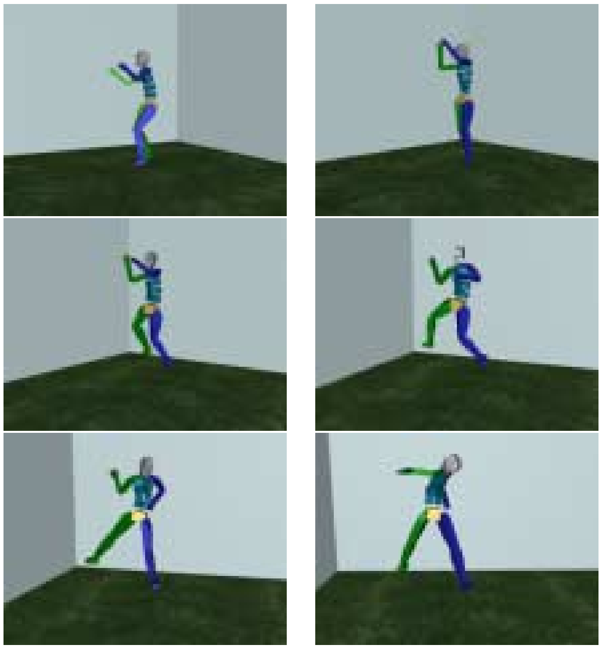


Figure 9: Six frames from an animation of the biped representation of the football example.

5. Conclusion

This paper has introduced a new constraint on the rectification of affine reconstructions from a known metric ratio of lengths of a pair of line segments. The constraint has been successfuly applied to motion capture by modelling a moving human as an articulated rigid link structure. Examples of reconstructed motion from uncalibrated broadcast footage of sports events have been presented, with both fixed and changing cameras. A number of extensions are possible.

As a system, the motion capture approach would benefit greatly from automated tracking to replace, or complement, the manual selection of joints. Completely automated tracking is unlikely to be completely successful given the quality of some digitised broadcast video. However, even partial tracking in conjunction with operator interaction will significantly speed up the process. The rigid link constraint could also be applied to marker data, where reflective markers are attached to the body to facilitate tracking.

A partial solution to the occlusion problem can be found when a section of semi-occluded frames can be omitted from a sequence, but the body can still be reconstructed from the remaining frames. An occluded point, now a *known* distance from an unoccluded 3D point, can lie anywhere on a sphere centred on the unoccluded point. If the occluded point is visible in one view, the back-projected ray defined by the image point intersects the sphere in two points, one of which is the occluded 3D point. Future work will incorporate this feature in the algorithm.

Finally, bundle adjustment is desirable to fit a truly rigid linked model to the image data. This is a complex non-linear optimisation problem, also currently under investigation.

References

[1] O. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *Proc. European Conference on Computer Vision*, LNCS 588, pages 563–578. Springer-Verlag, 1992.

[2] O. D. Faugeras. Stratification of three-dimensional vision: projective, affine, and metric representation. *Journal of the Optical Society of America*, A12:465–484, 1995.

[3] O. D. Faugeras and S. J. Maybank. Motion from point matches: Multiplicity of solutions. *International Journal of Computer Vision*, 4:225–246, 1990.

[4] R. I. Hartley. Estimation of relative camera positions for uncalibrated cameras. In *Proc. European Conference on Computer Vision*, LNCS 588, pages 579–587. Springer-Verlag, 1992.

[5] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.

[6] J. J. Koenderink and A. J. van Doorn. Affine structure from motion. *J. Opt. Soc. Am. A*, 8(2):377–385, 1991.

[7] E. Kreysig. *Advanced Engineering Mathematics*. John Wiley and Sons, New York, 1993.

[8] M. Leventon and W. Freeman. Bayesian estimation of 3-D human motion from an image sequence. Technical Report TR-98-06, MERL, 1998.

[9] D. Liebowitz and A. Zisserman. Metric rectification for perspective images of planes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 482–488, June 1998.

[10] D. Liebowitz and A. Zisserman. Combining scene and auto-calibration constraints. In *Proc. 7th International Conference on Computer Vision, Kerkyra, Greece*, September 1999.

[11] S. Maybank and O. D. Faugeras. A theory of self-calibration of a moving camera. *International Journal of Computer Vision*, 8(2):123–151, 1992.

[12] J. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision*. MIT Press, 1992.

[13] C. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. In *Proc. 3rd European Conference on Computer Vision, Stockholm*, volume 2, pages 97–108, 1994.

[14] M. Pollefeys, L. Van Gool, and A. Oosterlinck. The modulus constraint: a new constraint for self-calibration. In *Proc. International Conference on Pattern Recognition*, pages 31–42, 1996.

[15] L. Quan. Self-calibration of an affine camera from multiple views. *International Journal of Computer Vision*, 19(1):93–105, 1996.

[16] 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, 1996.

[17] L. Shapiro, A. Zisserman, and M. Brady. Motion from point matches using affine epipolar geometry. In *Proc. European Conference on Computer Vision*, LNCS 800/801. Springer-Verlag, 1994.

[18] L. S. Shapiro, A. Zisserman, and M. Brady. 3D motion recovery via affine epipolar geometry. *International Journal of Computer Vision*, 16(2):147–182, 1995.

[19] C. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2000.

[20] 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, November 1992.

[21] J. A. Webb and J. K. Aggarwal. Structure from motion of rigid and jointed object. *Artificial Intelligence*, 19:107–130, 1983.

[22] D. Weinshall and C. Tomasi. Linear and incremental acquisition of invariant shape models from image sequences. In *Proc. 4th International Conference on Computer Vision, Berlin*, pages 675–682, Los Alamitos, CA, 1993. IEEE Computer Society Press.