Using Skeleton-Based Tracking to Increase the Reliability of Optical Motion Capture

Lorna HERDA, Pascal FUA, Ralf PLÄNKERS, Ronan BOULIC and Daniel THALMANN
{Lorna.Herda, Pascal.Fua, Ralf.Plaenkers, Ronan.Boulic, Daniel.Thalmann}@epfl.ch
Computer Graphics Lab (LIG)
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
Tel: +41 21 693 52 48
Fax: +41 21 693 53 28

Abstract

Optical motion capture provides an impressive ability to replicate gestures. However, even with a highly professional system there are many instances where crucial markers are occluded or when the algorithm confuses the trajectory of one marker with that of another. This requires much editing work on the user's part before the complete animation is ready for use. In this paper, we present an approach to increasing the robustness of a motion capture system by using an anatomical human model. It includes a reasonably precise description of the skeleton's mobility and an approximated envelope. It allows us to accurately predict the 3–D location and visibility of markers, thus significantly increasing the robustness of the marker tracking and assignment, and drastically reducing—or even eliminating—the need for human intervention during the 3–D reconstruction process.

Keywords: Biomechanics, gait, skeleton, tracking, motion.

1 Introduction

Optical motion capture has in the past years become an increasingly helpful tool in the area of human motion science, typically providing valuable information for assessing orthopaedic pathologies. In fact, most optical systems are manufactured for medical applications (Menache, 1999). The number of companies providing systems targeting the areas of biomechanics, sports performance and gait analysis reflects their relevance to these domains. Complete systems are available from various providers, such as the SPICAtek's Digital Motion Analysis System, MotionAnalysis' Gait Analysis, Charnwood Dynamics' CODA, Vicon Motion Systems' Clinical Manager, BTS' GAITeliclinic, Ariel Dynamics' APAS/Gait, and many more.
Motion capture systems allow the collection of information for illustrating and analysing the dynamics of gait, by studying the characteristics of body limbs and joints during various motions, such as walking, running, limb raising, etc. Discrepancies with respect to standard gait indicate some type of dysfunction. The output enables the physician to detect orthopaedic disorders, and to guide the subsequent treatment. It also enables him to determine whether the disorder has been effectively corrected, after treatment and/or surgery.

The movement is acquired with at least two cameras surrounding the scene, and the markers' 3-D positions are calculated. Using these, the aim is to infer the co-ordinates of the body joints and study their trajectory over the captured sequence.

Several major difficulties arise in the process. First of all, the reconstruction of the markers in 3-D must be accurate – this does not present great difficulties in the case where markers are visible from at least two cameras, but in the presence of occlusions, markers may be lost. The markers then need to be identified in each sequence frame, yielding a 3-D trajectory for each. This process is known as tracking. The identification of these markers implies real difficulties, and even with a highly professional system, there are many instances where crucial markers are occluded or when the algorithm confuses the trajectory of one marker with that of another. In the case where markers become occluded, it grinds to a halt and requires user guidance. Confusions typically occur because these markers are attached to skin or tight clothes that have their own relative motion with respect to the underlying bone structure. Indeed, if the marker identification process expects constant inter-marker distances or marker-to-joint distances, problems are guaranteed to arise. Finally, once the markers have been identified, the skeleton needs to be correctly associated with the cloud of 3-D markers, i.e. positioned correctly within them, to reflect the subject's posture.
The shortcomings of the previously mentioned process (see Fig.1a) limit its applicability in a real-time context and drive up post-processing costs for non real-time applications. This is the issue that our proposed approach - depicted by Fig.1b - addresses. In most commercially available packages, the estimation of the markers' 3–D positions and the fit of the 3–D skeleton are decoupled, whereas we first compute a skeleton-and-marker model using a standardised set of motions. We then use it to resolve the ambiguities during the 3–D reconstruction process.

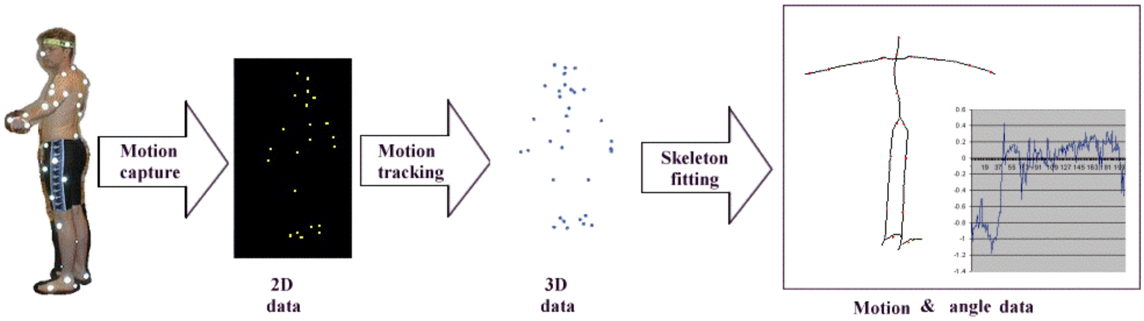

Fig.1a: Approaches to Motion Capture: Commercial packages

In this paper, we present an approach to increasing the robustness of a motion capture system by using a simplified anatomical human model. It includes a description of the skeleton's mobility and an approximated envelope. It allows us to predict the 3–D location and visibility of markers, thus significantly increasing the robustness of the marker tracking and assignment, and drastically reducing—or even eliminating— the need for human intervention during the 3–D reconstruction process. In contrast to commercially available approaches to motion capture such as the ones proposed by Elite^{TM} and VICON^{TM}, we do not treat 3–D marker reconstruction independently from motion recovery. Instead we combine these two processes and use prediction techniques to resolve ambiguities. For example, we can predict whether or not a marker is expected to be occluded by the body in one or more images and take this knowledge into account for reconstruction purposes. When a marker cannot be reconstructed with certainty from its image projections, we use the expected position of the skeleton to identify the marker and disambiguate its 3–D location. This is helpful when it is only seen by a small number of cameras. In our approach, the subject's skeleton motion is a by-product of the reconstruction process.

In the remainder of this paper, we present the theory behind 3-D marker reconstruction, tracking and identification. We then move on to describing the actual processing flow, applied to measured data, and then present the outcomes. We also provide some analysis of the system's performance, demonstrating its robustness using some complex motions that feature both large accelerations and severe occlusions.

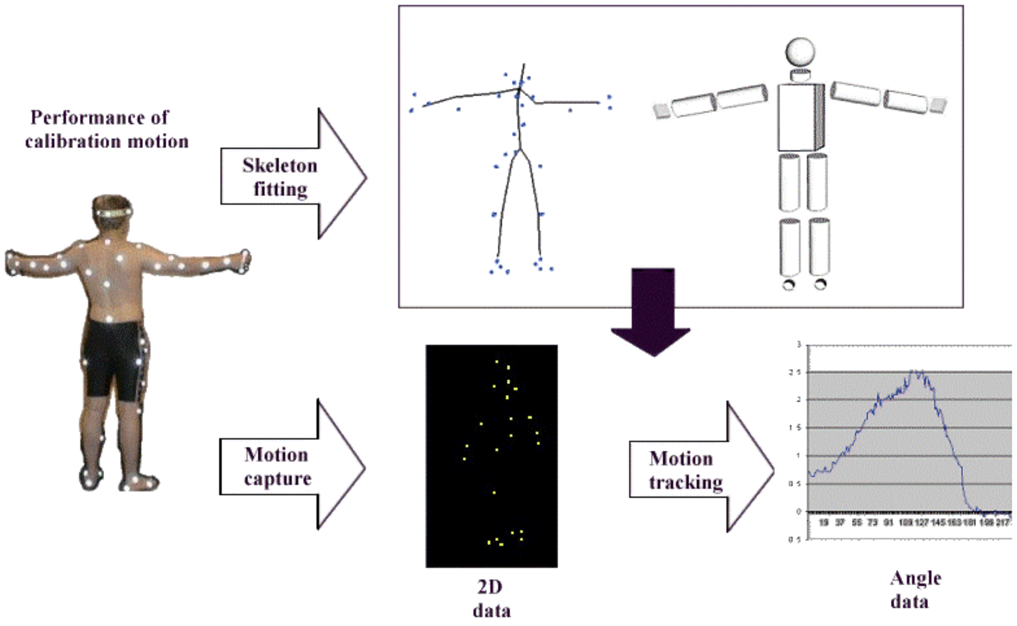

Fig.1b: Approaches to Motion Capture: Our approach.

2 Theory

The two major steps leading from a captured motion to a reconstructed one are:

- marker reconstruction from 2-D marker sets to 3-D positions;
- marker tracking from one frame to the next, in 2-D and/or 3-D.

However, despite the fact that 2–D and 3–D tracking ensure the identification of a large number of markers from one frame to another, ambiguities, sudden acceleration or occlusions will often cause erroneous reconstructions or breaks in the tracking links. For this reason, it has proved to be necessary to increase our procedure's robustness by using the skeleton to drive the reconstruction and tracking process by introducing a third step, i.e. the accurate identification of each 3-D marker and complete marker inventory in each frame.

The approaches to solving these issues are addressed in the following paragraphs, starting with the presentation of the human model used and keeping in mind that our entire approach is based on the constant interaction between the model and the above marker processing tasks.

2.1 Skeleton model

Our skeleton model is controlled by 32 degrees of freedom grouped in 15 joints, 14 bone lengths, and six position parameters in 3–D space. This is a simplified version of the complete skeleton generally used in our research lab for animating virtual humans. It does not include detailed hands and feet. The reasons for this modification are explained in section 3.2.2.

This model has a marker model associated to it, where the markers are attached to specific joints and are constrained to remain on a sphere centred in that joint. Obviously, this is a simplistic representation that does not reflect the effective relative motion of markers around joints. If more precision were necessary, this model should be refined, as will be discussed in Section 4.3.4.

2.2 3–D marker reconstruction

2.2.1 Stereo triangulation

3–D markers are reconstructed from the 2–D data using stereo triangulation (Faugeras and Robert, 1996). Given the projection P_1 of a point in an image, the corresponding projection in another image must lie on the epipolar line. The correspondence between the two images is established by the fundamental matrix, computed on the basis of the calibration data of the cameras.

In our case of eight camera views, we perform the test of the epipolar constraint on a set of two views, thus performing pair-wise reconstruction. For each non-ambiguous stereo match, that is when there is only one possible candidate, we compute the 3-D co-ordinates on the basis of the 2-D co-ordinates. The 2-D markers whose co-ordinates were used for reconstruction are assigned an associated 3-D index corresponding to the created 3-D marker.

These 3-D co-ordinates are then re-projected onto the remaining six camera views, to determine the corresponding 2-D markers. These 2-D markers found on re-projection are assigned a secondary 3-D index corresponding to the reconstructed 3-D marker. In this manner, for each 3-D marker, we can determine the complete set of corresponding 2-D markers in the camera views.

We assume that a 3-D marker is correctly reconstructed if it re-projects into at least one other camera view (thus making a total of three camera views). We will say that these markers are reconstructed by trinocular stereo, i.e. using at least three cameras. This is in contrast to markers reconstructed using only two camera views, and for which projections in the other views could not be found.

2.2.2 Binocular reconstruction

Once we have reconstructed these trinocular 3–D markers in the first frame, we need to compare the number of reconstructed markers with the number of markers known to be carried by the subject. As all remaining processing is automatic, it is absolutely essential that all markers be identified in the first frame. Any marker not present in the first frame is lost for the entire sequence. Therefore, if the number of reconstructed markers is insufficient, a second stereo matching is performed, this time also taking into account markers seen in only two views.

2.2.3 Testing the accuracy of the reconstruction

In order to improve the results of stereo matching, we use the skeleton for applying a visibility and occlusion test to each pair of 2–D markers used to construct a 3–D marker, thus verifying the validity of the reconstruction.

Visibility Check A marker is expected to be visible in a given view if it is seen more or less face on as opposed to edge on, that is if the surface normal at the marker's location and the line of sight form an acute angle. Suppose that we have reconstructed a certain 3–D marker using the 2–D pair (marker i_1, view j_1) and (marker i_2, view j_2); we check that these two markers i_1 and i_2 are indeed visible in views j_1 and j_2 respectively. Still assuming that displacement is minimal from one frame to the next, we use the skeleton's posture in the previous frame and calculate the normal at the 3–D marker's location with respect to its underlying body part segment. We draw the line joining the 3–D marker co-ordinates to the position in space of the camera and if the angle between the normal and the line is acute, then the marker is visible. If this test shows that we have used the wrong 2–D co-ordinates for reconstruction, we must select other candidate 2–D co-ordinates. As discussed in Section 2.2.1, each 3–D marker is associated to two sets of 2–D co-ordinates determined by stereo correspondence, which we then use for reconstructing the 3–D marker. To this 3–D marker, we then also associate the 2–D co-ordinates from the remaining camera views onto which the 3–D co-ordinates of the marker projected correctly. Given that the visibility test has detected an erroneous 3–D reconstruction, we choose one of the 2–D co-ordinates computed via 3–D to 2–D projection, and calculate new 3–D co-ordinates. We then perform a new visibility test, and if this fails, we repeat the entire procedure. The visibility test is illustrated by Fig.2a.

Occlusion check Once a 3–D marker has passed the visibility test, it needs to undergo the occlusion check: We want to ensure that the 3–D marker is not occluded from some camera views by another body part. To this end, we approximate body parts by solids, cylinders for limbs and a sphere for the head. In the case of limbs, the cylinder's axis is the corresponding bone and the radius is the average joint-to-marker distance of the markers associated to this body part. In the case of the sphere, the centre is the mid-point of the segment. For each 3–D marker, a line is traced from the marker to the position of the camera, and tested for intersection with all body part solids. In case of intersection with a solid, the marker is most likely occluded from this camera view (see Fig.2b). Therefore, we conclude that we have used erroneous 2–D co-ordinates for reconstruction. As before, we choose other 2–D co-ordinates and repeat the process.

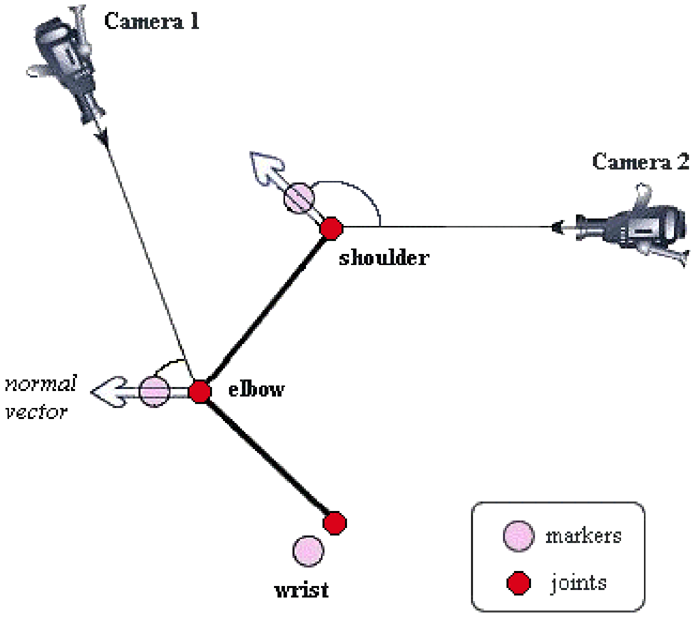

Fig.2a: Visibility test; the elbow marker is clearly visible to camera C_1, as the normal of the marker with respect to the skeleton produces an acute angle with the ray connecting the optical centre of the camera and the elbow joint. The shoulder joint, on the other hand, is not visible from camera C_2, as the angle formed is obtuse.

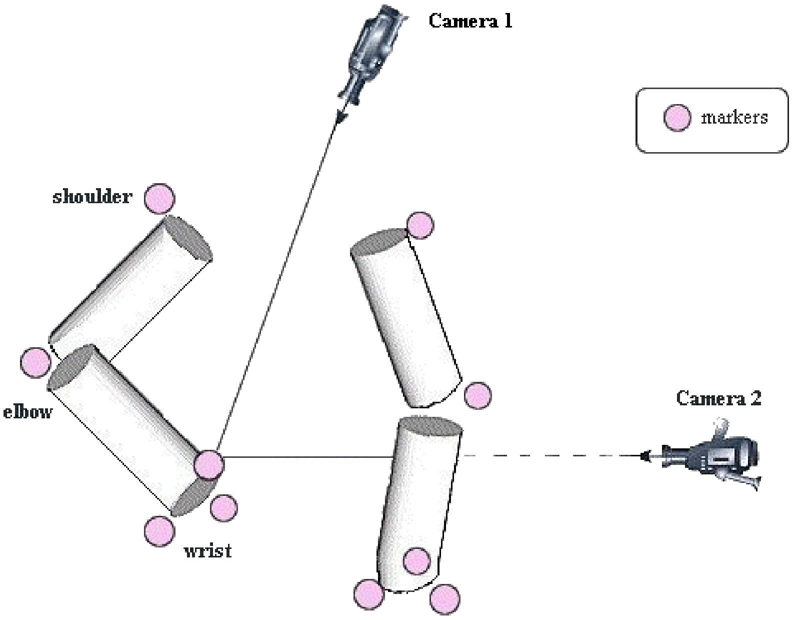

Fig.2b: Occlusion test; the marker on the left wrist is visible from camera C_1, as the line connecting the camera's optical centre and the wrist marker do not intersect any of the body solids. The marker is however occluded by the right lower arm body part, obstructing it from the view of camera C_2.

2.3 Tracking

Once as many markers as possible have been reconstructed, we can proceed with tracking of these markers from one frame into the next, thus resulting in marker trajectories over the entire sequence
2–D tracking is carried out at the same time as 3–D tracking because 2–D sequences are bound to provide more continuity than reconstructed 3–D sequences. We therefore use 2–D tracking in order to accelerate 3-D reconstruction: for each reliably reconstructed marker in frame [f], we consider the two sets of 2–D co-ordinates that were used to compute its 3–D co-ordinates. After 2–D tracking, these two sets of 2–D co-ordinates will most likely have links to two sets of 2–D co-ordinates in [f+1], the next frame. If so, we can then use them in [f+1] to construct the corresponding 3–D marker. To determine the related 2–D positions in the other camera views, we re-project the 3–D co-ordinates, as in the stereo matching process described above.

3–D tracking propagates the information attached to each marker in the first frame throughout the entire calibration motion, so that as many markers as possible are identified in all frames. A broken link in the tracked trajectory of a marker would imply the loss of its identity and a subsequent user intervention. In paragraph 2.4, we will see how we use the skeleton to overcome that problem in an automated fashion. To compute the trajectory of a marker from frame [f] into frame [f+1], both in 2–D and 3–D, we look at the displacement of the marker over a four-frame sliding window (Malik, Dracos and Papantoniou, 1993).

The basic assumption is that displacement is minimal from one frame into the next, and the idea is to predict and confirm the position of a marker in the next frame. The displacement of a marker from [f-1] into [f] predicts the position in [f+1]. The actual position in [f+1] and the projection of the movement into [f+2] should confirm the previously made hypothesis by eliminating ambiguities. In other words, we follow the displacement of markers over a sliding window of four frames.

Three cases can arise for each marker in [f] that we attempt to track into [f+1]:

- The marker has a tracking link into [f-1]: in this case, we calculate the displacement from [f-1] into [f] and apply it to the current position of the marker in [f], maintaining the direction of the movement. Once we have a predicted position, we define a *search neighbourhood* in [f+1] centred in the predicted position, this being the region in which we expect to effectively locate the marker. If within this neighbourhood we find more than one candidate, we prolong the movement into [f+2] in exactly the same fashion, for each candidate in [f+1]. The trajectory from [f-1] to [f+2] that has the smoothest acceleration will determine the effective track from [f] into [f+1].
- The marker has no tracking link; we define a *correlation neighbourhood* in [f], centred in the current position of the marker in [f]. This neighbourhood should include markers whose movement is correlated to the marker in focus. We use the links from [f-1] into [f] of those markers to determine the displacement to apply to the marker. We then proceed as above.
- There is neither of the above; we can only define an arbitrary displacement (based for example on the average displacement that we expect for markers from one frame to another), and proceed as above.

In all cases, the predicted position is not considered to be a fixed point, but a segment whose extremities are the predicted position assuming zero displacement and the predicted position calculated by one of the three above-mentioned schemes. We assume that the effective position will be on this segment. Using only the predicted position on the basis of acceleration introduced errors in the case where the subject suddenly remained immobile from one frame to another. Indeed, if we apply the acceleration observed in the *n* previous frames to a marker in the current frame, this will obviously yield a marker displaced with respect to its previous position. However, if the human subject is no longer moving but is immobile, the predicted position will be inaccurate and will produce errors in the reconstruction.

In the example of Fig.3, the marker in focus in [f] has a link in [f-1], on the basis of which we prolong the movement into [f+1]. In the search neighbourhood formed, there are two possible candidates. Prolonging the movement into [f+2] yields another two candidates. Of the two trajectories, we choose the one with the smoothest acceleration over the four frames, which then determines the marker in [f+1] that is indeed the correct link to the one in [f].

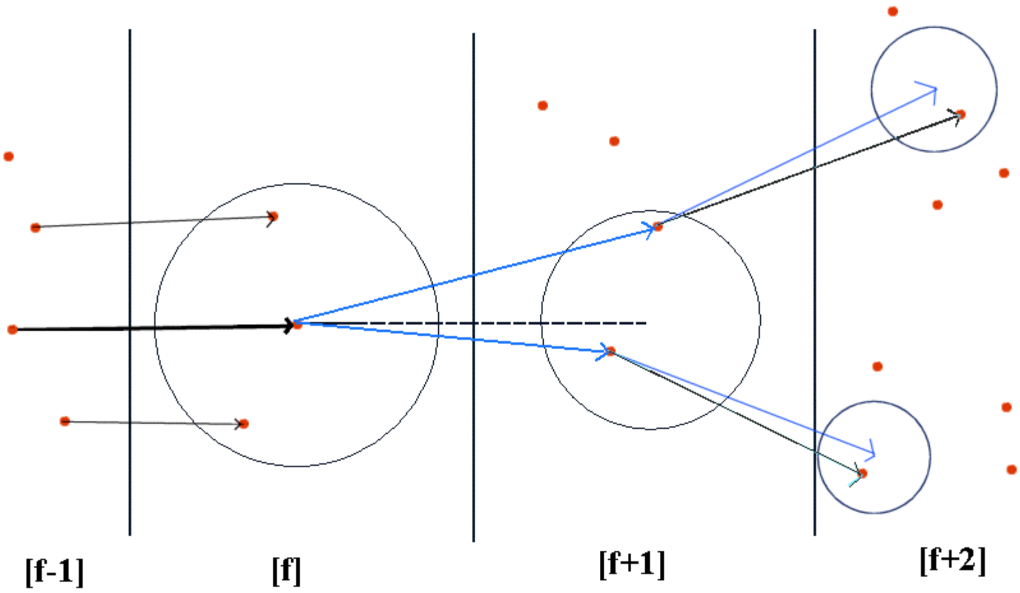

Fig.3: Illustration of the four-frame tracking principle, where [f] is the current frame that we are tracking into frame [f+1], [f-1] is the previous frame enabling us to construct a correlation neighbourhood, and [f+2] being the acceleration frame that solves eventual ambiguities among the tracking candidates.

2.4 Marker identification and inventory

The identification of all markers is necessary in order to be able to associate the skeleton model to the cloud of markers. However, tracking alone is mostly not sufficient for this purpose, due to tracking limitations, noise in the measured data, etc. Combining the model and marker tracking should allow us to complete the job, as well as to perform a marker inventory, as each marker needs to be present in each frame, and the position of a missing marker to be inferred.

Say we have just performed 3–D reconstruction using the 2–D data of frame [f], and we have thus obtained a set of markers. We then proceed with the following checks:

1. Binocular reconstruction: If the number of markers reconstructed using trinocular stereo is smaller than the actual number of markers worn by the subject, we perform binocular reconstruction and add the newly calculated co-ordinates to the already existing list of markers.

2. Identification by tracking: We perform 3–D tracking from [f-1] into [f], thus identifying a certain number of markers in [f], i.e. attaching them to their legitimate joint. 3–D tracking is performed in two passes, the first one being regular 3–D tracking with the predefined search neighbourhood, and the second pass attempts to track the still unlinked markers but allowing expansion of the search neighbourhood by a certain factor (to be set by the user). As the neighbourhood is statically set by the user upon initialisation, its value is not necessarily optimal for the entire sequence, as the speed of displacement may vary. Expanding the neighbourhood on a second 3–D tracking pass remedies this situation to a certain extent.

3. Monocular reconstruction: If the number of reconstructed markers is still smaller than the actual number of markers worn by the subject, we resort to monocular reconstruction. For this, we browse through the set of 2–D markers that are not associated to any reconstructed 3–D marker, and this for all cameras. For each such 2–D position, we calculate the optical centre of its associated camera, and calculate the equation of the line emanating from the camera and passing through the 2–D position, thus giving us a line in 3–D world space on which the corresponding 3–D marker should be located. For each marker that still remains to be localised in the current frame, we retrieve its position in the previous frame, as well as the position in the previous frame of its underlying joint. We then retrieve its marker-to-joint distance and create a sphere centred in the joint of radius marker-to-joint distance. If the line emanating from the optical centre intersects (at one or two points) with this sphere, then we are able to re-create the missing marker. In the case of one intersection, there is no ambiguity. In the case of two intersections, we will retain the co-ordinates that most closely match the position of the marker in the previous frame. In the case of monocular reconstruction, the reconstructed marker is already identified, we have therefore separated it from the trinocular and binocular reconstructions, as no tracking is necessary (Fig.4).

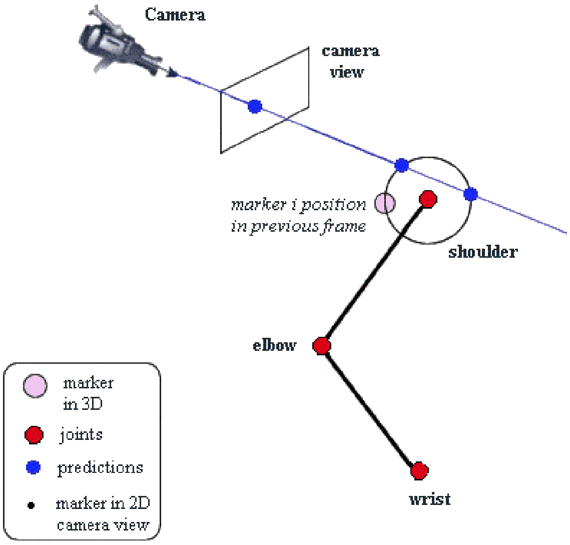

Fig.4a: One-camera marker reconstruction principle; the ray cast from the camera's optical centre through the image point of marker i intersects (in world space) the sphere around the joint associated to that marker. There are two intersection points, the one retained is the one whose position is closest to the marker's position in the previous 3D frame.

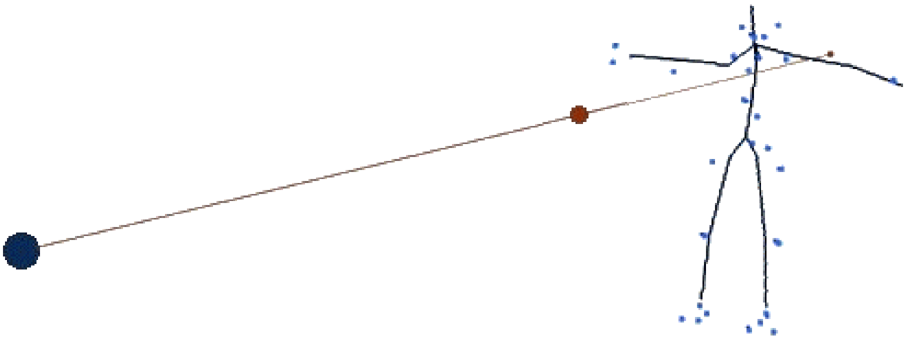

Fig.4b: Camera ray in 3D scene, yielding the reconstruction of an elbow marker.

4. Identification by closest-limb and likelihood estimation: If all markers are still not found, we attempt to identify the 3–D markers that are still anonymous. We find all the skeleton's joints that are missing one or more markers. Assuming that displacement is minimal from one frame to another, we retrieve the co-ordinates of these joints in the previous frame, and calculate the distance from these joints to each remaining unidentified 3–D marker; the distance closest to the marker-to-joint distance specified by the marker model yields an association of the 3–D marker to that joint.

We verify that this association is plausible by using 2-D tracking and 3-D tracking consistency. In other words, if a marker is supposedly identified as being marker *i* in the current frame [f], then it is linked in a trajectory to marker *i* in frame [f-1]. The set of 2-D markers associated to the 3-D markers, in [f-1] and [f], should logically also be linked in trajectories. We make sure that this is the case, as discrepancies would indicate that the association is erroneous. This is based on the fact that statistically, 2-D tracking is more trust-worthy than marker-to-joint association – namely due to the fact that 2-D positions are reliable and that there is continuity from one frame to the next.

Before accepting the association, a second test is performed, namely the segment rigidity test. The skeleton being an articulated structure with rigid segments, the distance between two joints on a same segment is logically always the same. Therefore, we can extend this fact to the markers associated with the joints. Given that the markers are positioned on a sphere around the joint, the distance from one marker to another is not fixed and rigid, but it is basically contained within an interval whose minimum value is the length of the underlying bone segment minus the two marker-to-joint distances, and whose maximum is the length of the segment plus the two distances (the minimum possible distance between the two markers is [bone_length-d1-d2] and the maximum distance is [bone_length+d1+d2] at the extremes of the spheres).

We check that these distances match for all markers on a same segment, for the associations found, and if there is a none-match, the association is rejected.

Combining 'elastic' segment lengths, flexible marker-to-joint distances and a closest-neighbour matching between markers and skeleton, we have ensured the robustness of the system with respect to artefact motion of the markers. Indeed, marker-to-marker and marker-to-joint distances can vary without causing a breakdown of the algorithm, due to the fact that these are not eliminatory conditions for marker identification.

5. Position correction: If the distance from marker to joint is larger than the distance specified by the marker model, we ”bind” the co-ordinates of the 3–D marker to the joint: We change its 3–D co-ordinates so that the marker moves within an acceptable distance of the joint. We however leave all reliably reconstructed 3–D markers untouched. This option can be enabled or disabled by the user, at will. This process is reliable if the variance of the measured marker-to-joint distances in the calibration phase was not too large and if sufficient data was provided in order to calculate the marker-to-joint distances.

In the case where after all these steps, we still do not have the complete number of markers, we need to estimate these and create them:

6. Marker position prediction: In the worst-case scenario, there may still be joints that are missing markers. We retrieve these markers in the three previous frames [f-3], [f-2] and [f-1], and calculate the acceleration; we apply this acceleration to the position in [f-1], thus obtaining an estimated position of the marker in the current frame [f]. As before, we calculate the distance from this inferred position to its associated joint. If it is out of range, we ”bind” the co-ordinates, the binding distance being a pre-set parameter multiplied by the marker-to-joint distance. In this manner, all 3–D markers are available for the fitting process.

3 Experimental validation

In the previous section, we introduced our approach, here we present the complete workflow from input to output, and the various steps involved.

3.1 Motion capture input data

We use as input the 2–D camera data and calibration parameters provided by an Elite TM optical motion capture system (Ferrigno and Pedotti, 1985). More precisely, the subject wears markers and is imaged by eight infrared cameras. The Elite TM system returns a 2–D location for each visible marker. We will hence be working on sets of 2–D point locations, one for each marker and each camera that sees it, and a projection matrix for each camera.

To extract a 3–D animation of a skeleton from a variety of movements performed by the same subject wearing the same markers, we first need to derive the skeleton-and-marker model, that is a skeleton scaled to the subject's body proportions and an estimate of the markers' locations with respect to the joints. To achieve this result, the subject is asked to perform a "calibration motion," that is, a sequence of simple movements that involve all the major body joints. We can then use this calibrated skeleton for further motion capture sessions of more complex motions.

The complete process is described in Fig.5.

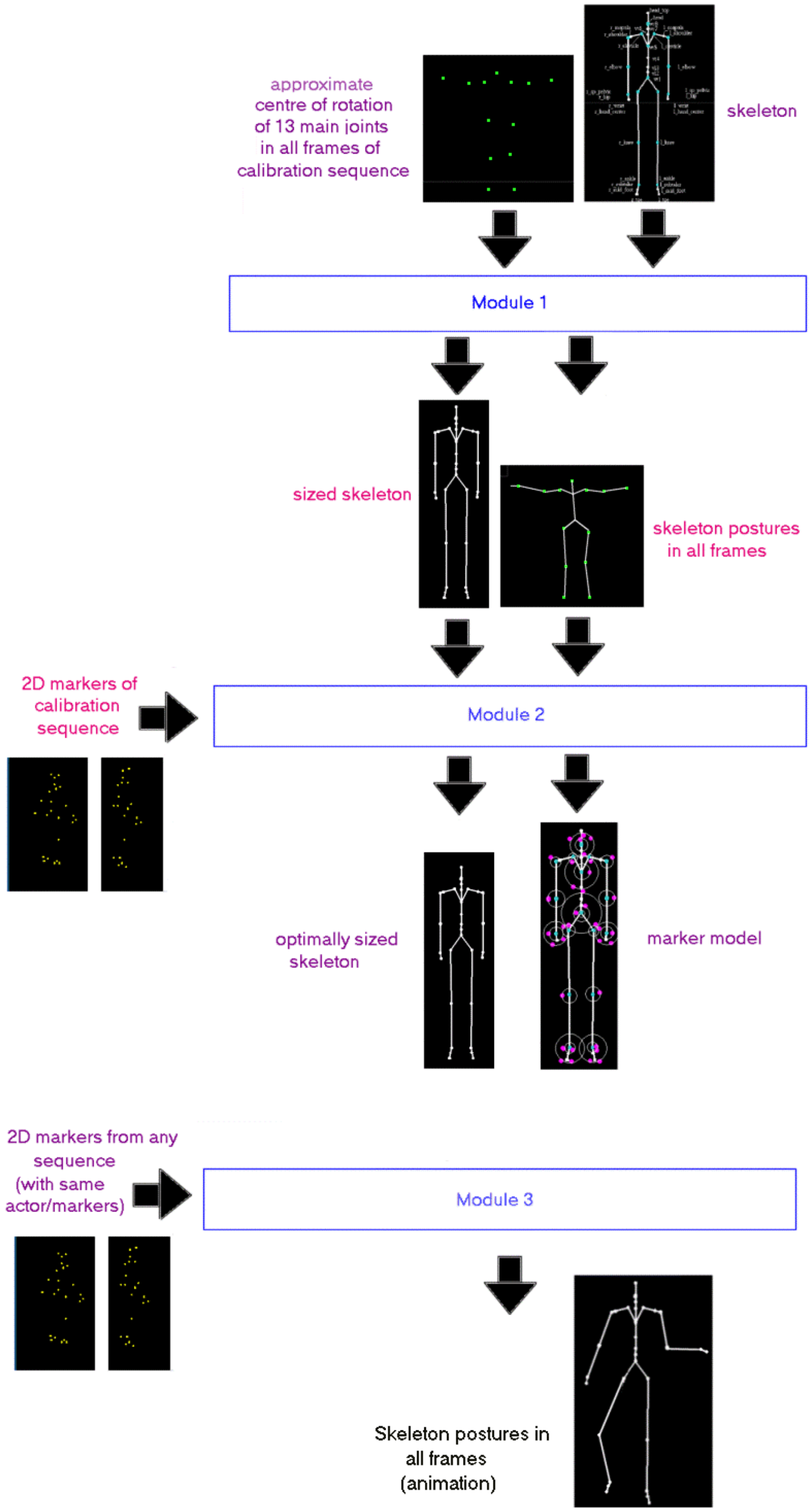

Fig. 5: Summary of skeleton-based tracking and fitting modules.

We need to start - as for each captured motion - by reconstructing the markers in 3-D. As the calibration motion is an especially simple routine highlighting the major joint motions, the 3–D location of the markers can be automatically and reliably reconstructed without knowledge of the skeleton for 200 to 300 successive frames. In practice, we partition the calibration motion into independent sequences, each one involving only the motion of one limb or body part at a time. We then perform 3–D reconstruction and tracking for each one independently. If necessary, the user can reattach some markers to specific body parts if they become lost.

In the first frame of the sequence, 3-D reconstruction using trinocular stereo matching is followed by a binocular pass, in order to attempt to reconstruct the missing markers. Binocular stereo matching is bound to introduce errors so at this stage, the user is prompted to confirm whether or not these binocular reconstructions are correct, as any error is bound to be propagated throughout the process.

As soon as all markers are found in the first frame, the user is asked to associate each marker to a joint. For each highlighted marker, the user must select a body part and corresponding joint. Any marker not associated to a body part is discarded during the fitting process. Once these associations have been manually created, we can proceed with 2–D and 3–D tracking of the markers, thus giving us the calibration motion reconstructed in 3–D, the trajectories of the markers throughout the sequence, as well as the identification of the markers with respect to the skeleton model.

3.2 Acquiring the Skeleton and Marker Model

During the calibration phase, our goal is to scale the bones of the generic skeleton so that it conforms to the subject's anatomy and to model the marker's locations with respect to the joints.

The skeleton-and-marker model is computed by least-squares minimisation. As this is a non-linear process, the system goes through three successive adjustment steps so as to move closer and closer to the solution at an acceptable cost while avoiding local minima. These steps are described below.

3.2. 1 Initial Joint Localisation

In earlier work (Silaghi, Plänkers, Boulic, Fua and Thalmann, 1998), we have developed a non-iterative technique that allows us to use these tracked markers to roughly estimate the 3–D location of a few key joints in each frame of the sequence, as well as the relative 3–D trajectories of the markers with respect to the underlying joints. We describe this technique briefly below and refer the interested reader to our earlier publication for additional details.

Let us consider a referential bound to a bone represented as a segment. Under the assumption that the distance between markers and joints remains constant, the markers that are attached on adjacent segments move on a sphere centred on the joint that links the two segments. The position of a segment in space is completely defined by three points. Thus, if we have a minimum of three markers on a segment, we can define the position and orientation of that segment in space. Afterwards, we compute the movement of the markers on adjacent segments in the referential established by these markers and we estimate their centres of rotation.

To take advantage of this observation, we partition the markers into sets that appear to move rigidly and estimate the 3–D location of the centre of rotation between adjacent subsets, which corresponds to the joint location.

This yields the approximate 3–D location of thirteen major joints, namely the joints of the arms and legs, as well as the location of the pelvic joint, at the base of the spine.

3.2.2 Skeleton Initialisation

Given these thirteen joint locations in all frames, we take the median distances between them to be estimates of the length of the subject's limbs. We then use anthropometric tables to infer the length of the other skeleton segments.

This gives us a skeleton model scaled to the size of the subject. This model, however, is a static one, meaning it has the appropriate dimensions but does not yet capture the postures for the calibration sequence or the relative position of markers and joints.

To estimate those marker-to-joint distances, we first need to roughly position the skeleton in each frame by minimising the distance of the thirteen key joints to the corresponding centres of rotation. This is done by minimising an objective function that is the sum of square distances from the centres of rotation to the joint it is attached to.

Given the fact that we use a sampling rate of 100 Hertz and that the calibration motion is slow, the displacement from one frame to another is very small. Fitting is performed one frame at a time, and the initial parameter values for frame [f] are the optimised parameters obtained from the fitting in the previous frame [f-1]. As we only have thirteen observations for each frame, we do not attempt to estimate all of the skeleton's degrees of freedom. Only ten joints (shoulders, elbows, hips, knees, pelvic joint and the fourth spine vertebra) are active while all the others remain frozen. This yields the postures of the skeleton in all frames of the calibration motion. In other words, we now have values of the global positioning vectors and degrees of freedom in each frame, as well as a better approximation to the limb lengths of the skeleton.

Least-squares sense fitting is applied using the joint positions of the local fitting as observations, and the skeleton as the fitting model, in other words, we will modify the parameters of the model in order to minimise the objective function, i.e. the distance **d** from the observations to the model.

The least-squares sense minimisation scheme stipulates that the function to be minimised is the sum of the squares of the errors. For a given frame, the error of the fitting is the current distance from the observation to the model. In the absence of noise, this distance should be zero.

The joint positions calculated by the local fitting module are in world-space co-ordinates. In order to measure the distance from the observation to the model, we transform the joint positions from global world co-ordinates to the model's local joint co-ordinates. Once this transformation has been applied, the distance is simply the norm of the observation (as the joint of the model is now the referential at co-ordinates (0, 0, 0)). The square sum of errors should be zero, as we are fitting approximated joint positions to the model's optimised joint positions.

The square sum of errors is: $\sum_{\text{first\_frame}}^{\text{last\_frame}} \left\| \text{obs\_local} \right\|^2 = 0$

where $\left\|\text{obs\_local}\right\| = \sqrt{x^2 + y^2 + z^2}$ is the objective function to minimise, (x, y, z) being the local joint co-ordinates of the observation.

For each frame, the error is:

$$\boxed{err = \sum_{first\_frame}^{last\_frame} (x_{shoulder}{}^2 + y_{shoulder}{}^2 + z_{shoulder}{}^2) + (x_{elbow}{}^2 + y_{elbow}{}^2 + z_{elbow}{}^2) + (x_{wrist}{}^2 + y_{wrist}{}^2 + z_{wrist}{}^2)}$$

for the example of the arm.

The parameters of the model that are to be adjusted are the following:

- the bone lengths
- the global position in space of the model (x, y, z of the translation vector, and τ, ρ, κ of the rotation vector)
- the degrees of freedom (DOF) of the joints, each joint having from one to three DOF.

In order for the least-square fitting to converge towards an optimal value for each parameter, the number of parameters to solve must not be too high with respect to the number of observations (between 30 and 40, typically), otherwise the system would be under-constrained. This is the reason why we have simplified the skeleton in order for the model parameters to be reduced to the following:

- 32 degrees of freedom (15 joints)
- 14 bone lengths (plus the lengths of neck, hand and feet segments)
- 6 parameters for setting the global position of the skeleton in 3-D space

Minimising the objective function means determining the zeros of its partial derivatives with respect to all model parameters. The derivatives that are calculated are those with respect to the parameters x, y and z, i.e. the position of the parent joint of the associated body part. The co-ordinates of the joint are themselves function of the global position and degrees of freedom of the joints preceding it in the hierarchy.

For each frame, and each observation, we calculate the derivatives of the objective function (one for each body part) and adjust the parameters of the model in order to minimise the objective function. This procedure is repeated a number of times, this number of iterations being pre-set. Usually, 3 to 5 iterations yield sufficiently accurate estimations.

As the joint positions obtained by local fitting are all identified, we know to which body part each observation belongs, and to which joint of the model it should be fitted, so from here on, we can run the least-squares fitting process. We thus obtain the posture of the skeleton in all frames of the calibration motion. These postures will then serve the purpose of initial values of the model parameters for the next fitting stage. The closer the initial postures are to the solution, the less iterations will be necessary, this being the purpose of calculating these skeleton postures.

3.3 Global Fitting

We now have a skeleton model that is scaled to the size of the performing subject, but we are still missing a complete marker model, meaning one that specifies where the markers are positioned on the subject's body and their distance to the joints they are attached to. This is computed by performing a second least-squares minimisation where the actual 3–D marker locations become the data to which we intend to fit the skeleton.

Markers are not located exactly on the joints and the marker-to-joint distances must be estimated. To this end, we superimpose the markers' 3–D co-ordinates with the previously computed skeleton postures. In each frame, we then compute the distance from the marker to the joint and we take the median value of these distances to be our initial estimate of the marker-to-joint distance. This is the value of the objective function (sum of square errors) that we would like to attain. The principle is illustrated in Fig.6.

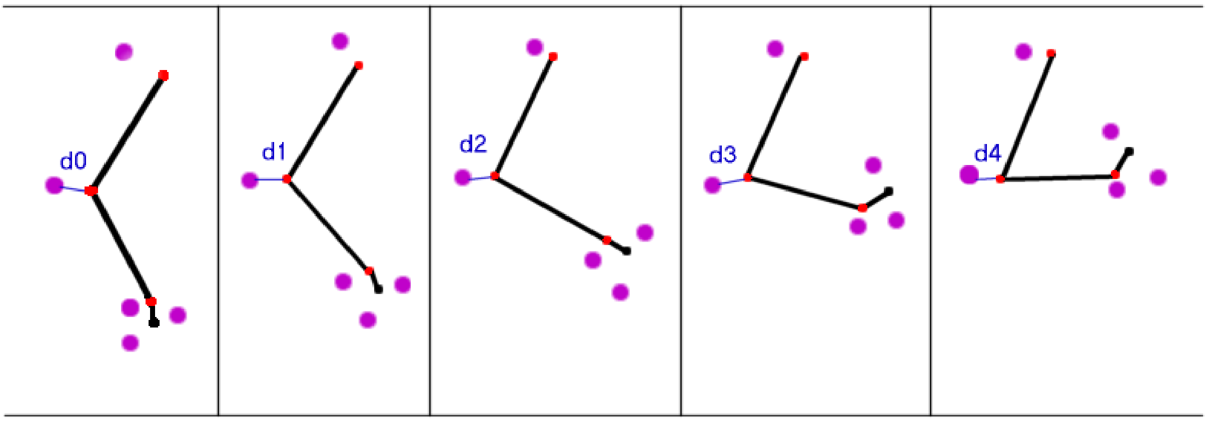

Fig.6: Initial median marker-to-joint distance estimate; for each of the five frames, we have the reconstructed

$$\sum_{first\_frame}^{last\_frame} \left\| obs\_local \right\|^{2} = \bar{d}$$

where obs_local are the co-ordinates of the observation in the local joint referential and
d is the median observation-to-joint distance

Taking the marker model to be the distance from marker to joint means that the marker is expected to always be located on a sphere centred in the joint.

We now have all the information required to fit the skeleton model to the observation data. The initial state is given by the previously obtained skeleton postures, and all markers are identified by the user in frame 1 through the interface of the application.

As we need to check that all markers are present and identified before fitting, we do it one frame at a time. Technically, the fitting process is similar to the one we used to fit models to stereo video sequences (Plänkers, Fua and D'Apuzzo, 1999). The interested reader is referred to (D'Apuzzo, Plänkers, Fua, Gruen and Thalmann, 1999) for details on the algorithm.
For each frame and for each marker, once the fitting is complete, the distance between marker and joint is stored. At the end of the calibration motion sequence, we have as many such distances per marker as there are frames. The median value of these distances is an improved approximation of the marker-to-joint distance and becomes the final marker model.

3.4 Capturing Complex Motions

The resulting skeleton-and-marker model can now be applied to motions that we actually wish to capture. The procedure is very similar to the one used in the global fitting step of the previous section, with the difference that the user is now required to identify the reconstructed 3-D markers in the first frame by associating them directly to 3–D markers located on the skeleton-and-marker model, as per the model computed during the calibration phase.

4 Outcomes

4.1 Improvements provided by the skeleton-based tracking

Here we attempt to quantify the improvements brought about by using the skeleton with respect to marker reconstruction and tracking.
In Fig.7, we present some statistics on the improvement resulting from the visibility and occlusion test with respect to accurate marker reconstruction. We used 400 frames of, respectively, the calibration and karate motion. We show the percentage of marker pairs that were rejected/accepted by the visibility/occlusion test. These statistics underline the fact that one third of the stereo correspondences were doubtful, and a better match was to be sought.
In Fig.8, we compare the number of correct stereo matching associations found by the algorithm with or without the visibility/occlusion test. Our reference for this test was a copy of the same 2-D file used for input, but with the markers having previously been manually identified (file provided by the motion capture studio. Note that the test produces a higher percentage of correct stereo matches than simple reconstruction. The percentage gained may seem small, but we need to keep in mind the fact that errors are propagated through 3-D tracking and will therefore affect the entire process.

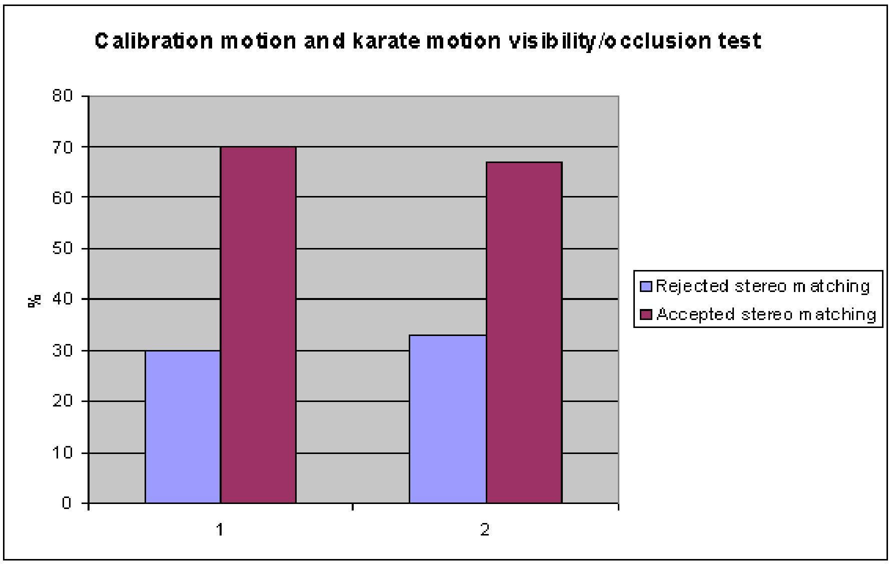

Fig.7: Visibility and occlusion test statistics over 400 frames

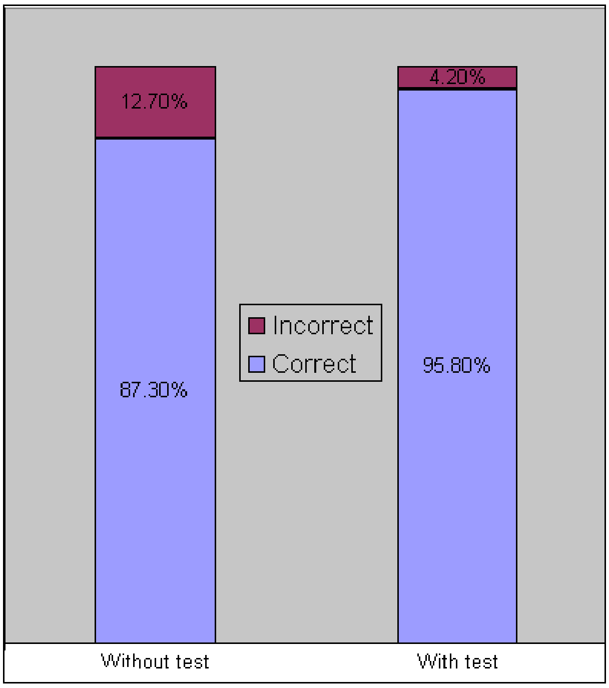

Fig.8: Visibility and occlusion test statistics over 800 frames of the gym motion, one trial without the test, and one with

In Fig.9, we have run 3-D tracking over 800 frames of the calibration motion. This tracking uses only simple marker prediction without skeleton information (see paragraph 3.1.1) over a sliding window of four frames. If a marker is lost in a frame, it cannot be recovered in any of the subsequent frames. The figure shows that after about 300 frames, the number of tracked markers drops to about 10% (tests performed with a subject using 32 body markers). In Fig.10, we have performed the same statistics, but this time using a captured motion containing a fast movement, and we notice that simple tracking loses all markers in less than 200 frames. Note that when we say that a marker is 'lost', this means that a complete trajectory from frame zero to the present frame is no longer available. The marker can still be tracked to a marker in the previous or following frames, but the trajectory will be fractured there where the marker was not identified.

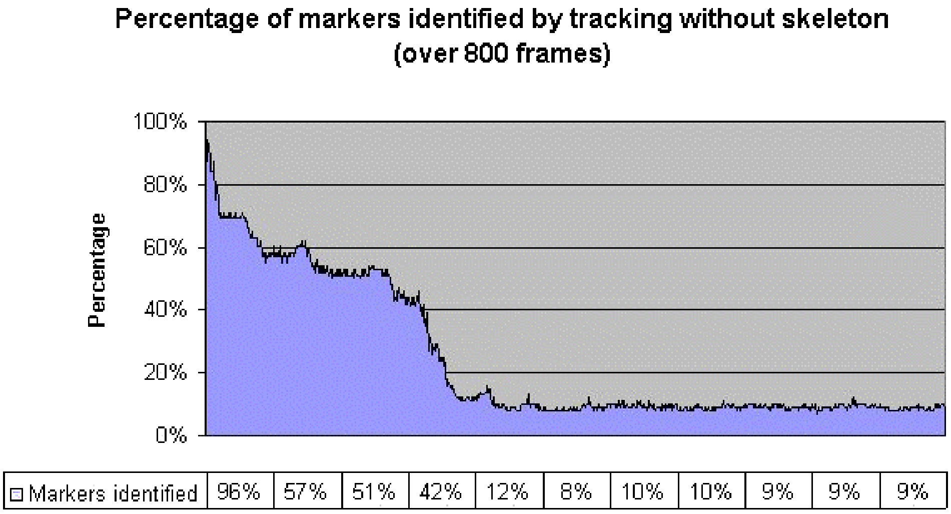

Fig.9: Percentage of markers identified by simple tracking, for the calibration motion.

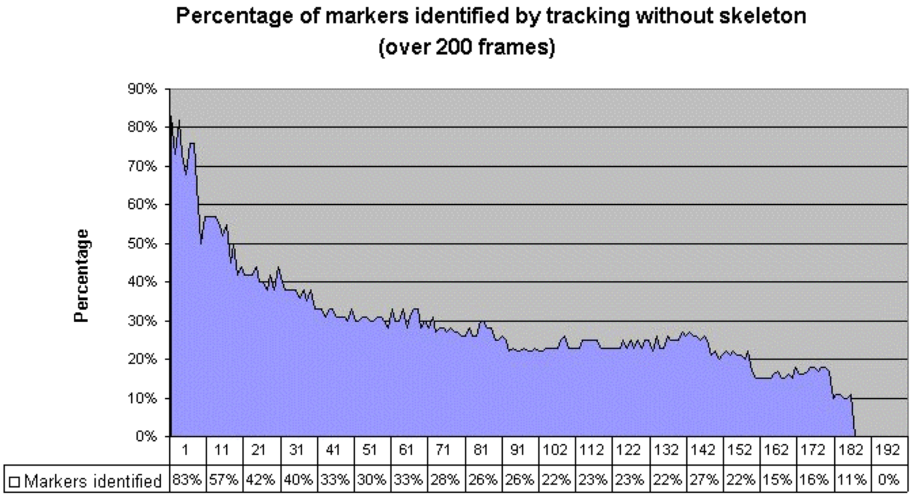

Fig.10: Percentage of markers identified by simple tracking, for a captured karate motion.

With skeleton-based tracking, all markers are recovered in all frames, i.e. 100% of the markers are present and identified – keeping in mind the fact that a certain percentage of these markers were reconstructed by prediction, their co-ordinates thus being subject to a certain error factor.

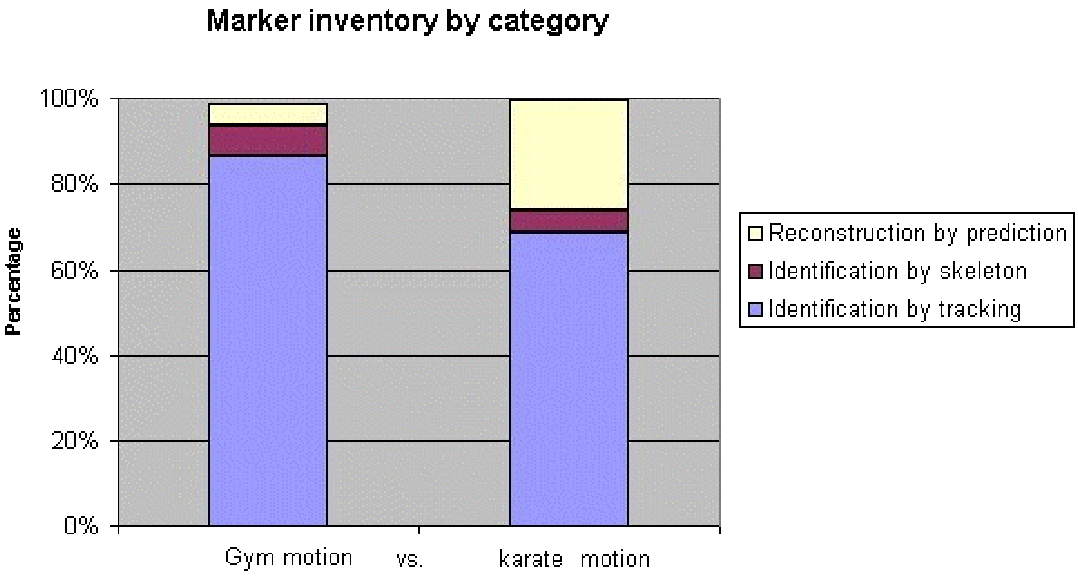

Fig.11: Percentage of markers identified by each step of the marker inventory, respectively for the calibration and the karate motion.

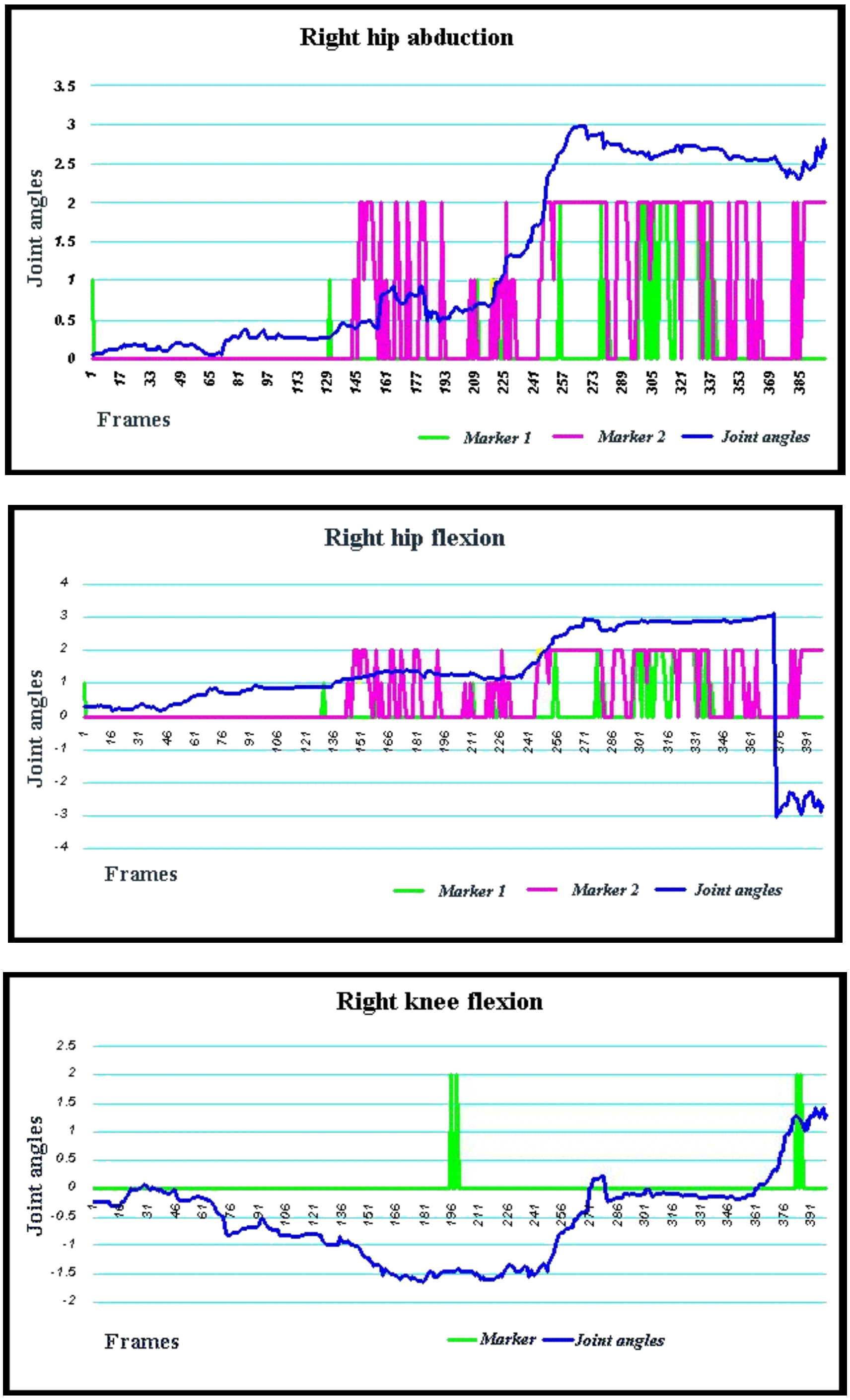

Fig.12: Graphs of the hip (a,b) and knee (c) DOF values over 400 frames, including statistics on marker identification

In Fig.11, we show the percentage of markers identified by each process of the marker inventory of skeleton-based tracking (see paragraph 2.4). The two bar charts compare these percentages in the case of the calibration motion and the karate motion. The first step of the process identifies reconstructed markers using simple 3-D tracking. This percentage is obviously higher than in the previous figures, because in the skeleton-based tracking case, all markers are present in each frame, whereas in the simple tracking case, a marker lost in one frame is lost forever. The second step identifies reconstructed markers using the position of the skeleton in the previous frame. As to the third step, it reconstructs the markers that are still missing in the current frame, combining prediction of the 3-D tracking type and the position of the underlying skeleton.

The obvious observation would be that identification by skeleton proximity and reconstruction by marker prediction have little influence, with respect to the performance of tracking. However, the reality is not that straightforward, if we keep in mind that a broken tracking link can never be recovered, no matter how efficient the tracking algorithm is. Therefore, full marker reconstruction and accurate identification are absolutely essential; as shown by Fig.9 and 10, the identification rate would soon drop to zero without it.

In Fig.12, we superimpose the right leg joint angle values onto the marker reconstruction and identification statistics so as to be able to determine whether any particular errors are introduced by these methods.

The peaks represent integer values of respectively 0, 1 or 2, in the cases where the markers are respectively identified through "*Identification by tracking*" (case 2, paragraph 3.2.2.*), "Identification by closest-limb and likelihood estimation*" (case 4), and "*Marker position prediction*" (case 6), for 400 frames of the karate motion. The figures show that following a marker on the knee is relatively easy and that tracking can cope with that single-handedly. However, for the more complex case of the hip joints, simple tracking is insufficient after 100 frames, and the alternative methods take over, without introducing gross errors.

4.2 Fitting Results

The images in Fig.13 and 14 show a few results obtained with one of the movements provided by the Motion Capture Studio. The use of the skeleton has enabled us to improve every step of the process, from 3-D reconstruction, to tracking and identification of the markers. It is robust with respect to noisy data, out-of-bound and non-identified markers will be rejected, and it can also handle the case of occluded markers.

4.3 Discussion

Today, most optical motion capture software systems require human intervention for solving ambiguous stereo matches, as well as for re-identifying markers when a broken tracking link occurs. In our case, the entire process is automatic, as soon as initialisation has been performed by the user.

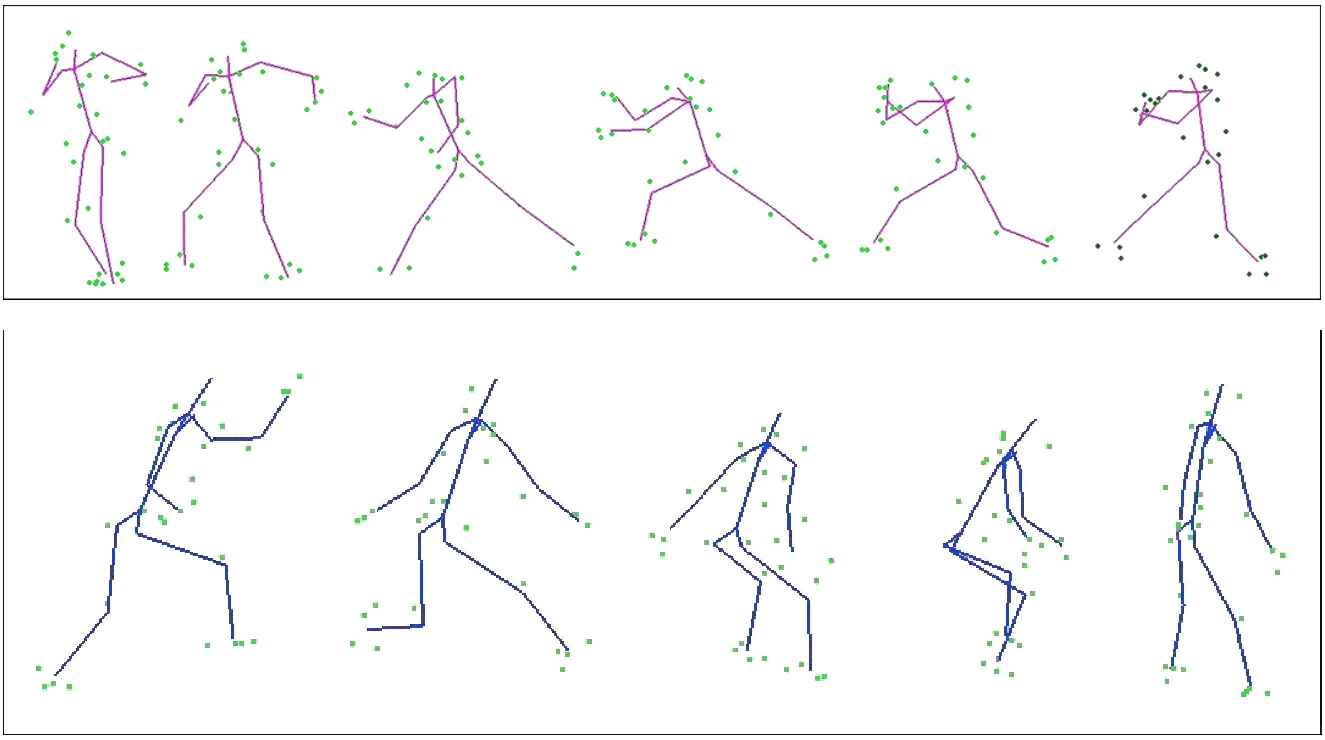

Fig.13: Karate and jump motions

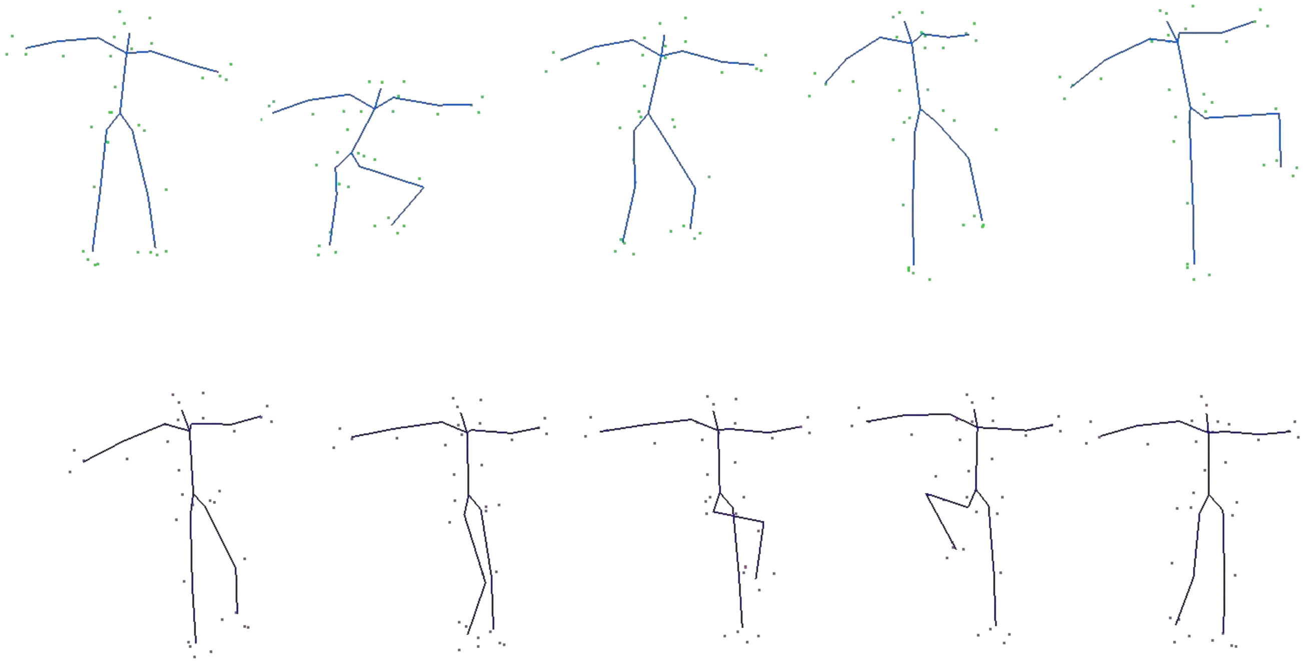

Fig.14: Knee and hip joint motions

As is the case of all methods, there is however space for improvement.

4.3.1 Pre-set thresholds and parameters

For tracking and identification, we could allow dynamic search neighbourhoods, which would render the algorithm more robust with respect to sudden accelerations. These neighbourhoods would be function of the acceleration, thus expanding their radius when the movement accelerates. The thresholds set for tracking and fitting are directly linked to the average displacement between two frames. Therefore, thresholds set at the beginning of the session in view of a regular movement at a certain speed are most often not optimal in case of a sudden acceleration.

The markers will move out of the boundaries of the search neighbourhoods defined by these thresholds, this resulting either in rejection of these markers (in the tracking case) or in these markers' co-ordinates being modified so as to move them back closer to the skeleton joints (in the identification case). For example, a rapid movement involving an extension will result in an extension that is not as full as in the original motion.

4.3.2 User intervention

With respect to initialisation, one could do without user intervention if the subject were to adopt a specific pose at the beginning of the sequence this being the norm for calibration in the context of motion capture.

4.3.3 Over-determination of the problem

Sometimes, there is no single solution to the fitting problem. This happens for example with the spine. Even when we keep only two joints in the spine, multiple solutions are possible for each posture, resulting in a twist at the spine level, and torsion or roll at the pelvic joint level. To solve this problem, we need to introduce some constraints on the spine on our side, or perhaps add more markers to the back of the subject. There are presently four markers on the torso: one at the bottom of the back, one at the level of the third vertebrae, one at the base of the neck, and one on the chest. This is obviously too few in order to determine a unique solution.

4.3.4 Marker model limitations

Regarding the marker model we are using for the moment, we could in the future use a more sophisticated model that would take into account the relative trajectories of the markers, calculated in the joint localisation phase (see paragraph 3.1.2). The markers are presently free to evolve on a sphere centred in the joint, whereas in reality, their relative movement is much more limited. This would enable us to be more precise when it comes to identifying a marker for sure, and also for reconstructing a missing marker around a joint.

4.5 Conclusion

We have presented an approach for increasing the robustness of an optical motion tracking device. We use a body-model during the 3–D reconstruction and tracking process to assist the 3–D reconstruction of the markers, take visibility constraints into account and remove ambiguities. This greatly increases the motion capture system's robustness and decreases the need for human intervention. We hope to further improve the precision of the reconstruction movement through future work that would incorporate biomechanical constraints and a more sophisticated skeleton.

The results shown in this paper were obtained using as input the data produced by a specific optical system. However, as we only use the markers' 2–D image locations and the camera calibration parameters, the approach is generic and could be incorporated in any similar system.

Acknowledgements:
The authors would like to thank Actisystem S.A. for providing the motion capture data.

References

[1] Badler N., Hollick M., & Granieri J. (1993). Real-Time Control of a virtual Human Using Minimal Sensors. *Presence*, 2(1):1–5.

[2] Chang, W. L., Su, F. C., Wu, H. W. & Wong, C. Y. (1998). Motion Analysis of Scapula with Combined Skeleton and Skin-based Marker System, *3rd World Congress of Biomechanics*, Sapporo, Japan.

[3] D'Apuzzo N., Plänkers R., Fua P., Gruen A. & Thalmann D. (1999). Modeling Human Bodies from Video Sequences, *Electronic Imaging*, *The International Society for Optical Engineering's Photonics West Symposium*, San Jose, CA.

[4] Faugeras O.D. & Robert L. (1996). What can two images tell us about a third one?, *International Journal of Computer Vision*, (18):5–19.

[5] Ferrigno G. & Pedotti A. (1985). Elite: A digital dedicate hardware system for movement analysis via real-time tv signal processing, *The Institute of Electrical and Electronics Engineers' Transactions on Biomedical Engineering*, BME-32(11).

[6] Halvorsen K., Lundberg A. & Lesser M. (1999). A new method for estimating the axis of rotation and the center of rotation, *Journal of Biomechanics*, Vol: 32, Issue: 11.

[7] Malik N., Dracos T. & Papantoniou D. (1993). Particle tracking in three-dimensional turbulent flows - Part II: Particle tracking, *Experiments in Fluids*, 15:279–294.

[8] Menache A. (1999). *Understanding Motion Capture for Computer Animation and Video Games*, Morgan Kaufmann Publishers.

[9] Plänkers R., Fua P. (2001). Articulated Soft Objects for Video-based Body Modeling, *International Conference on Computer Vision*, Vancouver, Canada.

[10] Silaghi M-C., Plänkers R., Boulic R., Fua P. & Thalmann D. (1998). Local and global skeleton fitting techniques for optical motion capture, *Workshop on Modelling and Motion Capture Techniques for Virtual Environments*, Geneva, Switzerland.