3D Reconstruction of Sports Events for Digital TV

Cornelius Malerczyk
ZGDV e.V., Computer Graphics Center
Fraunhoferstr. 5
64283, Darmstadt, Germany
cmalerc@zgdv.de

Konrad Klein
Fraunhofer Institute for Computer Graphics
Fraunhoferstr. 5
64283, Darmstadt, Germany
kklein@igd.fhg.de

Torsten Wiebesiek
Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung (TNT)
Appelstr. 9A
30167, Hannover, Germany
wiebesiek@tnt.uni-hannover.de

ABSTRACT

As the capabilities of video standards and receiver hardware are increasing towards integrated 3d animations, generating realistic content is now becoming a limiting factor. In this paper we present a new technique of generating 3d content from reality, i.e. from video sequences acquired with normal TV cameras. The major aim is to provide the TV viewer with animated 3d reconstructions of athletic events in MPEG-4 over Digital Video Broadcast (DVB), which allows for an immersive experience via free navigation and interaction on the receiver side. As intervention in the actual scene, e.g. by markers, is often prohibited, markerless computer vision techniques are used on the images from normal broadcasting cameras for the accurate estimation of an athlete's movements. The paper focuses on the key components for the realistic reconstruction of 3d geometric features, which are the calibration of moving TV cameras and the modelling of the moving athlete in its environment.

Keywords

3d Reconstruction, Digital Video Broadcast, Camera Calibration, Motion Estimation, Computer Vision

1. INTRODUCTION

In the last years, several methods of enhancement were introduced in sports television, e.g. a moving line enabling the comparison of an athlete's attempt with the world record, or the overlay of two competitors for comparison of their technique, e.g. in skiing. Due to the nature of ordinary television, these enhancements were previously limited to 2d sequences the TV viewer cannot interact with. With the advent of MPEG-4, advanced set-top boxes enable the interactive visualization of animated 3d content. However, the creation of suitable content that makes use of the 3d features of the MPEG-4 format is much more difficult than the production of ordinary TV content, particularly in case of 3d content representing actual real world events.

In order to bridge the gap between the technical possibilities of MPEG-4 and the tools available for creating high quality content, we aim at automatically converting ordinary images from TV cameras to a 3d scene description which contains an animated body model of the athlete in its 3d environment with accurate body movements. These 3d animations enable several novel viewing modalities:

- The TV viewer interactively specifies the position and direction of the camera while watching the sports event.
- Multiple athletes can be watched in parallel within the same environment in order to compare their attempts.
- By overlaying a metric grid, the athlete's attempt can be analysed in detail.

The work presented here is embedded in the European project PISTE, which covers the end-to-end chain for creation, transmission and reception of enhanced content during sports broadcasts close to the actual event. Along with the 3d reconstruction of moving athletes [Klein 02], PISTE also provides tools for a number of 2d enhancements as well as an authoring tool that allows the efficient administration of the content creation process and enables fast dynamic generation of content using templates [Walczak 02]. Moreover, the transmission over DVB

Journal of WSCG, Vol.11, No.1., ISSN 1213-6972
WSCG'2003, February 3-7, 2003, Plzen, Czech Republic.

and the development of a set-top box capable of displaying MPEG-4 streams is addressed within PISTE. This paper focuses on the work towards the 3d reconstruction of sports events.

2. RELATED WORK

The analysis and tracking of human motion has been an active area of computer vision research during the last years. Detailed surveys of several 2d and 3d computer vision based techniques and their applications can be found in [Moeslund 01] and [Gavrila 99]. Here only a few examples of markerless motion capturing applications shall be stated, which are most closely related to our approach. Chu et al [Chu 03] aim at reconstructing human motion by capturing human volumes from multiple images without any underlying model. This approach seems to be limited to poses, which can be recognised unambiguously. Sidenbladh et al [Sidenbladh 92] address the problem of 3d human motion tracking by representing 3d human motion with a large database of example motions. A probabilistic approach using a PCA-based dimensional reduction solves the problem of efficient search in a large training set. Lee et al [Lee 02] use an articulated human body model consisting of tapered 3d cones representing 14 body segments, which are projected onto 2d images and matched with extracted body silhouettes. A combination of analytical and synthesis-based methods is used for tracking human motions, which still needs a learned scene background for detection purposes. In this paper we propose an approach that is able to reconstruct a complete sequence of a sports trial solely by analysing the original video footage delivered by the TV broadcaster cameras using a sports specific kinematic model.

3. CHALLENGES IN COMPUTER VISION

In order to accurately reconstruct 3d movements of an athlete, four major problems have to be addressed: the separation of the athlete's moving limbs from the background, the calibration of the video frames, the estimation of the 3d pose and position, and the tracking of the overall movement in the sequence. The work in PISTE focuses on methods used to deal with these problems in the context of fast camera movements (causing motion blurring) and swifter, higher, and stronger action of the athletes. The latter is to be recovered in the Body Animation Parameters of MPEG-4. Additionally, a 3d model of the environment has to be reconstructed and aligned with the reconstructed athlete.

Previous results were often based on massive employment of manual techniques, i.e. the reconstruction is performed for each field of the video sequence separately by mouse-clicking some known features in the background as well as all relevant joint positions of the athletes. Carrying out this approach for an event with one athlete covered by two cameras requires as much as 2400 accurate mouse clicks per second of video footage. Alternatively, hardware sensors on the TV cameras can be used to track the camera movement (pan, tilt, and zoom) through the sequence, but they still require an estimation of the camera's relative orientation. This is often difficult due to insufficient overlap of the background shown in the camera images.

PISTE pursues the minimisation of user interaction by assuming some characteristics of the problem:

- Additional photographs are used to ensure sufficient mutual overlap of background shown in the images for calibration. These additional photographs are used at the same time for the 3d reconstruction of the environment.
- The TV cameras typically vary their orientation in pan, tilt, and zoom only, while their positions are constant, i.e. the camera position has to be estimated only once and not for every single video field. The remaining parameters can be estimated using 2d imaging techniques.
- The athlete possesses the typical shape and behaviour for a specific kind of sports. Consequently, a spatio-temporal model of the athlete's movements is used to evaluate and to predict the pose and position through the sequence.

Together with an integration of camera calibration, 3d modelling, and texturing, these approaches reduce the user interaction to a minimum.

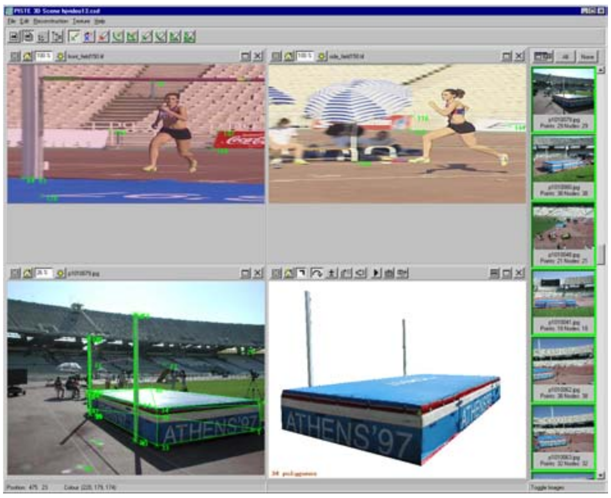

Figure 1. **3d reconstruction from multiple images.**

4. 3D RECONSTRUCTION

The 3d model of the environment is reconstructed from multiple photographs using well-known photogrammetric techniques. The primary goal is to reduce the number of required photographs, as their acquisition is often difficult in a sports TV context. To meet this demand, a wide-baseline approach has been chosen so that manual user interaction is required to identify corresponding features for camera calibration. Besides the standard approach based on 2d-2d correspondences, a 2d-3d approach, and a 3d-3d registration approach are used in order to reconstruct the scene from multiple photographic views [Hartley 00, Faugeras 93]. Subsequently, this camera set-up is optimised simultaneously using a bundle adjustment approach [Triggs 00].

As a very sparse point set represents the 3d surface at this point, we decided to model the surface by manually selected polygons rather than automatically increase the surface point density. This manual approach leads to small and efficient meshes (**Figure 1**).

In order to obtain a photorealistic result, the surface colour is determined from all photographs and sampled in OpenGL-compatible textures. The blending function used to combine the colours from the source photographs consists of four weights, which relate to visibility, quality (in terms of resolution on the 3d surface), a smooth transition to undefined regions, and a re-weighting of outlier colours using an M-Estimate approach [Neugeb 99].

This approach does not require any manual masking of persons in the scene as the respective regions are assigned a very low weight due to their outlier nature.

However, video footage has to be processed in the calibration step in order to include the broadcasting cameras in the reconstruction process. Fixed camera positions are still the common setup in many kinds of sport. An analysis of the camera setups in the Olympic games 2000 showed that the two required cameras are nearly always available. By assuming a constant camera position for the broadcasting cameras, the first-order primitive, which the calibration is based on, is not a single image, but rather an already stitched panorama consisting of a video sequence [Coorg 98]. The resulting advantage is twofold: the reduced number of unknown parameters eases in the calibration process, and the large field of view of the panorama increases the mathematical stability of the geometric set-up. The simultaneous estimation of camera parameters and 3d scene features is performed using a bundle block adjustment approach [Triggs 00]. Automatic early optimisation of parameters during the reconstruction ensures a good initial estimation of the overall optimisation problem, so that the processing time of the final optimisation is reduced to a few minutes.

Enriched with semantic information that can be generated within the same tool, the model is ready to be combined with an animated model of the athlete. The alignment of the athlete's model with the 3d environment is given without additional computation because the TV cameras that are used to estimate the athlete's 3d pose are calibrated according to the same coordinate system as the photographs that are used to calculate the 3d geometry of the environment.

5. RE-CALIBRATION OF BROADCASTING CAMERAS

The camera set-up consists not only of the static cameras used for the 3d reconstruction of the scene, but the non-static broadcasting cameras as well, which need to be re-calibrated when moved. We assume the position of the broadcasting cameras to be constant, so that the plenoptic function [McMillan 95] of each camera can be modelled by a panorama. However, to use this panorama image for calibration purposes as described in the next paragraph, it is necessary to relate this panorama image to the 3d camera set-up of the scene. To this end, we include a small number of video frames from each camera in the multi-camera calibration step. Subsequently, we use these images as "anchor points" within the panorama image of the respective camera so that both extrinsic and intrinsic parameters of the panorama image are known, namely the exact focal length and the camera position.

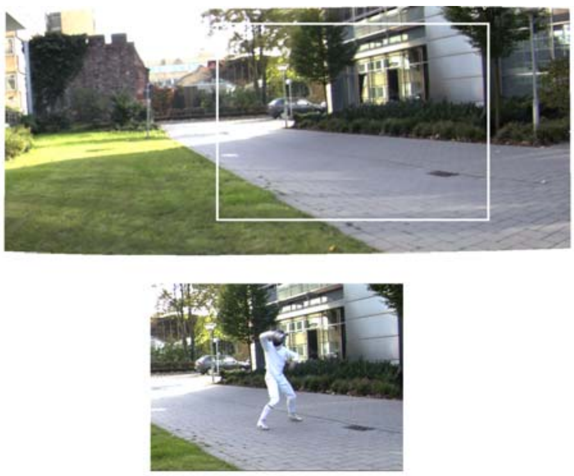

Figure 2. **Registration of single video fields with pre-calibrated environment map.**

The video footage showing the human body to be reconstructed can now be calibrated using 2d image registration techniques with the pre-computed panorama image (**Figure 2**). However, when comprehending the panorama image as the camera's plenoptic function, we ignored the time dimension. Consequently, the registration algorithm must be robust in spite of small changes in lighting conditions and moving persons in the scene, namely the athlete to be reconstructed.

To meet this demand, we use a hierarchical optimisation approach based on colour differences in the LUV colour space combined with an outlier re-weighting using M-Estimates [Zhang 97]. As the details of this 2d registration techniques are beyond the scope of this paper, we refer the reader to the literature on full view panorama stitching, e.g. [Szelski 97], where a similar problem is addressed.

Once a video frame is registered with the panorama image, its complete set of calibration parameters is known. The main advantage of this approach is the independency of the calibration of any other images except the panorama image. This allows camera positions that do not provide any overlap in the background scene, which is a usual condition in sports TV. Moreover, parallelisation can be employed as each single video frame can be processed independently.

6. 3D POSE ESTIMATION

In order to reconstruct an athlete's movements, synchronized and calibrated video sequences from at least two views are necessary. The calibration is carried out as described in the previous section. Both, the fast motion of the cameras and the athlete cause motion blurring. Additionally, we have to overcome difficulties introduced by self-occlusion. Therefore, we use a particular statistical model for each discipline that allows reliable temporal prediction of an athlete's pose. Moreover, we use an articulated 3d body model to exploit knowledge about human anatomy.

The images of each sequence are processed by a chain, which incorporates a number of computer vision techniques (**Figure 3**). At first, the athlete's silhouette is determined in each view by a seeded region growing algorithm [Sifakis 01, Adams 94]. Then an initial 3d pose is adapted to these observations. Therefore, a 3d body model is moved into the respective pose and projected into each view. Differences between segmented and synthetically created silhouettes are evaluated in order to determine the pose, which explains the observations best. From the 3d joint positions of the adapted body model, rotations for the joint angles are calculated to derive both VRML97 animations and MPEG-4 body animation parameter (BAP) used to animate an avatar at the receiver side. To overcome measurement errors like e.g. flickering of the athlete's movements, smoothing splines are used to reduce dithering effects within the completed animations.

In order to perform all these steps iteratively, automatically, and reliably, the initial pose is obtained by a prediction from previous poses. The prediction is based on a discipline specific, statistical model. This model is also able to detect a pose untypical for the specific kind of sports as an outlier, which requires confirmation or correction.

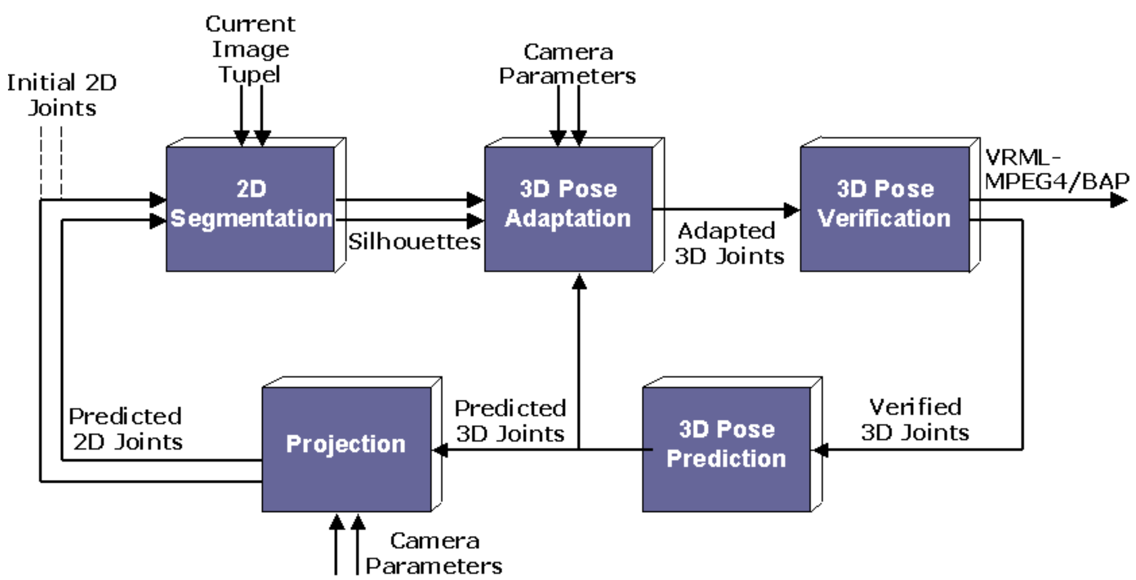

Figure 3: 3d pose estimation overview

The following section describes the pose prediction; the consecutive section explains the pose adaptation.

Pose Prediction

Within the computer vision pipeline, the 3d pose estimation as well as the motion prediction is needed for the correct representation of the athlete's body and its movements. In the PISTE project the human body is described as a set of 18 single joints, each representing a 3d position in the world coordinate system. Once the parameters of the TV cameras are known, two corresponding joint positions in image space are sufficient to determine the respective 3d position.

The kinematic information is calculated for each type of sports separately. This is done by a Point Distribution Model (PDM) [Cootes 92, Heap 96] of all possible poses of an athlete for a specific type of sports. The Point Distribution Model is a powerful shape description technique that may be used to derive a statistical description of objects from a set of training data. It is most useful for describing features that have well understood "general" shape, but which cannot be described by a rigid model. The human body is a good example for such a shape, that a human can comprehend and describe easily, but which do not permit rigid model-based description.

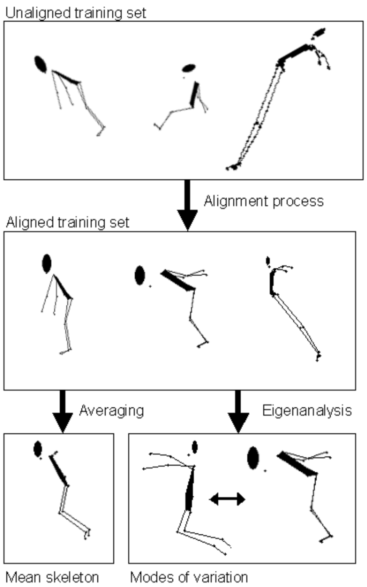

Figure 4. **Creation process of a PDM.**

In order to derive the statistical parameters from the training set, it is first necessary to align a set of 3d skeletons in an approximate sense (**Figure 4**). The minimization problem of the transformation function is an iterative application of a least-squares approach and can be solved by applying the Levenberg-Marquardt-Method [Marquardt 63].

The outcomes of this alignment process are (mutually aligned) 3d skeletons, from which it is possible to derive statistical parameters like the mean skeleton and the modes of variation. The knowledge of the mean skeleton allows explicit measurement of the variation and co-variation exhibited by each joint coordinate. Doing this for each aligned skeleton, we can calculate the covariance matrix, which has some useful properties. It exhibits the variations that are seen in the underlying training data.

These variations are important properties of the skeleton we are describing. The importance can be derived by an eigen-decomposition of the covariance matrix, which provides its eigenvectors and the eigenvalues. The eigenvectors associated with large eigenvalues correspond to large variation in the training data set. They provide the modes of variation (**Figure 5**).

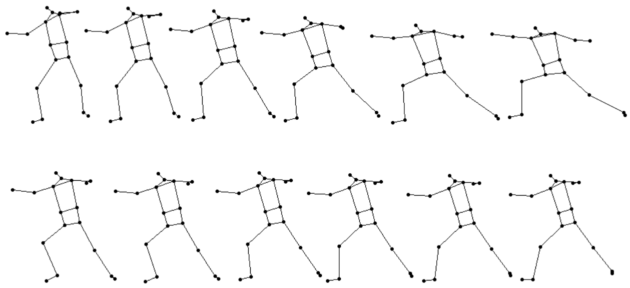

Figure 5. **1st and 2nd mode of variation for epee fencing.**

Sorting the eigenvalues by its numerical quantity, it is possible to represent any skeleton s (describes as a 54-dimensional vector of joint coordinates) as a linear combination of all eigenvectors:

$$s \approx \bar{s} + P_t b_t$$

where $\bar{s}$ is the mean skeleton, P_t the matrix of eigenvectors corresponding to the t largest eigenvalues and b_t the deformation vector. The components of this deformation vector b_t indicate how much variation is exhibited with respect to each of the corresponding eigenvectors in P_t.

Within the modes of variation the PDM detects a pose very untypical for the specific kind of sports as an outlier, which requires confirmation or correction by resolving the above equation into

$$b_t \approx P_t^{-1} \cdot (s - \bar{s})$$

Due to the fact that P_t is because of the dimensional reduction factor t not a square matrix, the calculation of the pseudo inverse matrix of P_t is needed to determine the corresponding deformation vector.

Verification of an arbitrary skeleton is now possible using the above equation by applying statistical limits to all components of the deformation vector. Poses are accepted as valid ones, if all components come to lie within this limiting interval.

The PDM can also be used to perform a pose correction automatically, e.g. if a limb is invisible or ambiguous in all images and the most reasonable pose must be found instead, while additional user interaction is requested as last resort only. Mathematically, mapping outliers of the deformation vector to its nearest valid values leads to the correction of an invalid skeleton.

With the knowledge of n last fields in a given sequence, it is possible to predict a pose in the following field by applying a non-linear extrapolation on the PDM´s parameters, more precisely by applying the extrapolation on the most important components of the deformation vector b_t, and calculating a new skeleton s using the above mentioned equation. This predicted 3d skeleton than is projected to the image planes and used as the next seed in the segmentation module.

Pose Adaptation

Aiming at an accurate description of the individual motion of an athlete during a particular attempt, the predicted pose has to be adapted to the actual observations. Therefore, we use silhouette information from multiple views. Silhouettes are obtained by a segmentation based on a Seeded Region Growing approach applied to the input camera images. Due to motion blurring and camera calibration errors, the segmentation results might be insufficient for some fields. A robust adaptation approach is required to overcome such a problem.

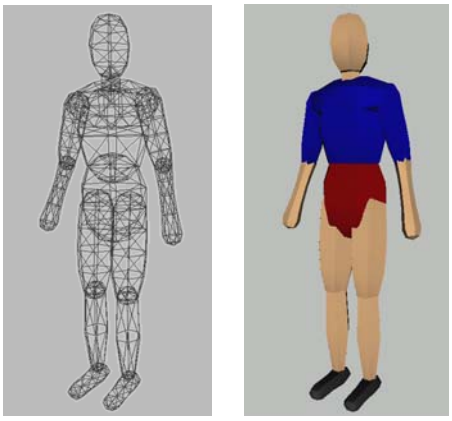

Figure 6. Wireframe and shaded representation of 3d body model.

We propose an analysis by synthesis approach, which can be subdivided into three major steps. Firstly, a generic 3d body model is set into the predicted pose. Then synthetic silhouettes of this model are generated for each available view by a fast rendering procedure using the results of the online camera calibration. In the third step, differences between synthetic and segmented silhouettes are analysed. These steps are repeated varying the pose hierarchically until an optimal explanation of the observed silhouettes is achieved.

The generic 3d body model consists of 15 simple volumetric primitives, which are attached to an articulated skeleton, represented by 18 joints (**Figure 6**). Approximate body proportions are taken from anthropometrical descriptions of human bodies like e.g. [Dreyfuss 67]. The number of polygons, which represent a body part, is kept low. So the model is simple enough for fast rendering of synthetic views, and complex enough to capture the pose of an athlete well.

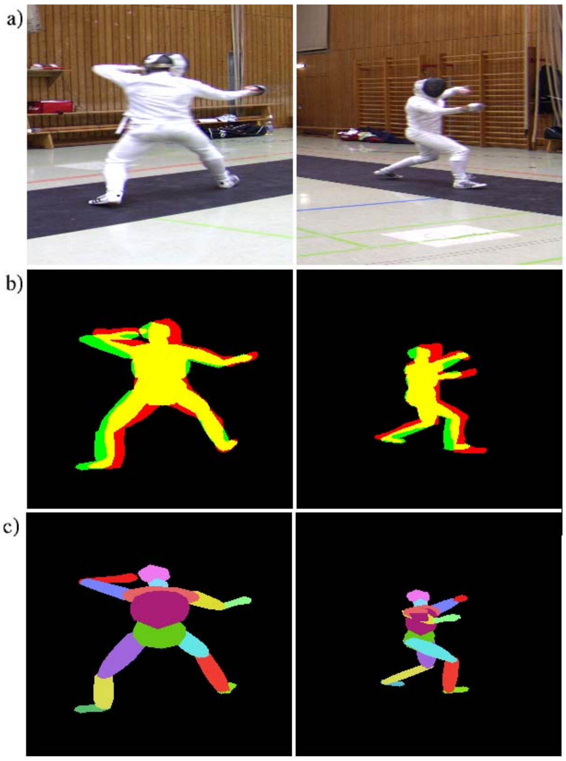

Figure 7. Input images (a), overlays of synthetic and observed silhouettes before pose adaptation (b), and label images (c).

In order to compare the pose of the 3d model and the real pose of the athlete, the 3d model is rendered into the image plane of each camera. For efficient rendering a simple pinhole camera model is used. Differences between the synthetic and observed silhouettes are evaluated in order to measure the correctness of the current pose. Figure 7a shows two views of a fencing sequence; in figure 7b corresponding overlay images of observed and synthetic silhouette can be seen. Here, green areas indicate parts of the observed silhouette, which are not covered by the synthetic ones, and red areas indicate the vice versa situation. Both types areas indicate an erroneous pose. In order to detect and overcome self-occlusion problems, not only synthetic silhouettes of the 3d model but label images are used. Each label in figure 7c represents one of the 15 body parts mentioned above. By using such label images, partial and complete occlusions of body parts can be recognized.

The 3d pose adaptation is performed hierarchically. At first, the general position and orientation of torso, chest and belly are adapted by fitting the corresponding body parts to the observed silhouettes. Figure 8a shows the initial configuration of these body parts. Let L_v be the ratio of label area outside the observed silhouette of view v (green) to the complete label area (green and yellow), and let S_v be the ratio of uncovered silhouette area in view v (red) to the complete silhouette area (red and yellow). The pose adaptation is carried out by varying 3d position and orientation of these body parts in order to minimize the sum of these ratios over all available views. To equally distribute the silhouette error over the body part contour, a third term is considered. C_v is the distance between the centers of gravity of the green and yellow areas in each view. The empirically determined weighting factors w_L, w_S and w_C ensure comparability between these terms.

This leads to the global error function *E*, which has to be minimized over all available views:

$$E = \sum_{v=1}^{V} w_L L_v + w_S S_v + w_C C_v$$

Figure 8b shows torso, chest and belly after adaptation. The pose adaptation is carried out by varying 3d position and orientation of each body part in order to minimize this error criterion by Powell's minimization strategy [Press 92]. Subsequently, the body limbs are optimized. Final adaptation results are shown in figure 9.

Figure 8. **Overlay images before (a) and after (b) torso, chest and belly adaptation.**

Figure 9**: Overlay images before (a) and after (b) adaptation.**

7. CONCLUSION

The MPEG-4 standard enables interactive, immersive TV experience on advanced set-top boxes, but the creation of suitable content that represents real world events is a significant bottleneck. The PISTE project addresses this challenge by developing content creation tools, which enable extensively automated 3d reconstruction from real world camera images (**Figure 10**). Even in cases where the TV camera set-up alone does not provide enough information for 3d computations, the integration of additional photographs leads to accurate results. Thus, the creation of 3d content encoded in MPEG-4 is made possible within one hour after the event, which will significantly advance interactive 3d television and increase its attractiveness.

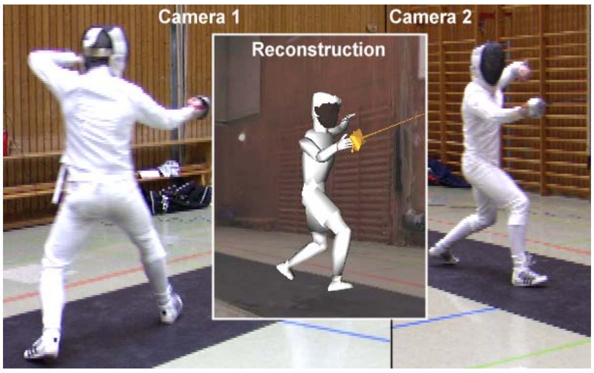

Figure 10**: Final reconstruction of epee fencing**

8. ACKNOWLEDGEMENT

The work presented here was partly funded by the European Commission within the IST project PISTE ("Personalised, Immersive Sports TV Experience"), contract number IST-1999-11172.

9. REFERENCES

[Adams 94] R. Adams, L. Bischof, “Seeded Region Growing”, IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-16:641-647, June 1994.

[Chu 03] C. Chu, O. Jenkins, M. Mataric. Towards Model-free Markerless Motion Capture, To appear in IEEE Intl. Conf. On Robotics and Automation, Taipei, Taiwan, 2003

[Coorg 98] S. Coorg, N. Master, S. Teller. Acquisition of a Large Pose-Mosaic Dataset, Proc. CVPR 1998, pp 872-878.

[Cootes 92] T. Cootes, C. Taylor, D. Cooper and J Graham. “Training Models of Shape from Sets of Examples”. Springer Verlag, London, 1992

[Demiris 01] A. Demiris, M. Traka, E. Reusens, K. Walczak, C. Garcia, K. Klein, C. Malerczyk, P. Kerbiriou, C. Bouville, E. Boyle, N. Ioannidis. Enhanced Sports Broadcasting by Means of Augmented Reality in MPEG-4, International Conference on Augmented, Virtual Environments and Three-Dimensional Imaging, Mykonos, Greece, 2001.

[Dreyfuss 67] “The Measure of Man: Human Factors in Design, 2nd edition”, Henry Dreyfuss & Associates, 1967.

[Faugeras 93] O. Faugeras. Three Dimensional Computer Vision - a Geometric Viewpoint, AI MIT Press, Cambridge MA, 1993.

[Gavrila 99] D. Gavrila. The Visual Analysis of Human Movement: A Survey, Computer Vision and Image Understanding, Vol 73, 1999.

[Hartley 00] R. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision, Cambridge Univ. press, 2000.

[Heap 96] T. Heap, D. Hogg. “Towards 3D Hand Tracking using a Deformable Model”. International conference on automatic face and gesture recognition, Oct 1996, Killington, Vermont, IEEE Computer Society Press, Los Alamitos, 1996

[Klein 02] K. Klein, C. Malerczyk, T. Wiebesiek, J. Wingbermühle, Creating a „Personalised, Immersive Sports TV Experience via 3d Reconstruction of Moving Athletes”, In W. Abramowicz (ed.), Business Information Systems, Proceedings of BIS, Poznan, 2002

[Lee 02] M. Lee, I. Cohen, S. Jung. Particle Filter with Analytical Inference for Human Body Tracking, IEEE Workshop on Motion and Video Computing, 2002.

[Marquardt 63] D.W. Marquardt, Journal of the Society for Industrial and Applied Mathematics, vol. 11, pp. 431–441, 1963

[McMillan 95] L. McMillan and G. Bishop, Plenoptic Modeling: An Image-Based Rendering System, Computer Graphics vol. 29, pp.39-46, 1995

[Moeslund 01] T. Moeslund and E. Granum, A Survey of Computer Vision-Based Human Motion Capture, Computer Vision and Image Understanding no. 81, pp.231-268, 2001

[Neugeb 99] Peter J. Neugebauer and Konrad Klein, “Texturing 3d Models of Real World Objects from Multiple Unregistered Photographic Views”. In Eurographics '99 Proceedings, Milan, September 7-11, 1999.

[Press 92] William H. Press et al, “Numerical recipes in C – The art of scientific Computing, 2nd edition”, Cambridge University Press, 1992.

[Sidenbladh 92] H. Sidenbladh, M. Black, L. Sigal, Implicit Probabilistic Models of Human Motion for Synthesis and Tracking, European Conference on Computer Vision, ECCV, 2002.

[Sifakis 01] E. Sifakis, C. Garcia, G. Tziritas, “Bayesian Level Sets for Image Segmentation”, Journal of Visual Communication and Image Representation, 2001.

[Szeliski 97] R. Szeliski and H. Shum, Creating full view panoramic image mosaics and environment maps, Computer Graphics vol.31, pp 251--258, 1997.

[Triggs 00] B. Triggs, P. McLauchlan, R. Hartley, A. Fitzgibbon. Bundle Adjustment - A Modern Synthesis, in *Vision Algorithms: Theory and Practice*, editor: Triggs, W. and Zisserman, A. and Szeliski, R., Springer LNCS 1883, pp 298-375, 2000

[Walczak 02] K. Walczak. X-VRML – XML Based Modeling of Virtual Reality. SAINT-2002 Proceedings, 2002.

[Zhang 97] Z. Zhang, Parameter Estimation Techniques: Tutorial with Application to Conic Fitting, International Journal of Image and Vision Computing, Vol.15, No.1, pages 59-76, 1997