Creating a "Personalised, Immersive Sports TV Experience" via 3d Reconstruction of Moving Athletes

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

Cornelius Malerczyk
ZGDV Computer Graphics Center, Darmstadt, Germany
cmalerc@zgdv.de

Torsten Wiebesiek, Jochen Wingbermühle
Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung (TNT)
[wiebesiek,wingber]@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.

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. Along with the 3d reconstruction of moving athletes, 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 [Walczak02]. Moreover, the transmission over DVB 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. 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 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.

3. 3d Reconstruction

The 3d model of the environment is reconstructed from multiple photographs using photogrammetric techniques. In our approach, the user is solely required to identify the corners of the objects to be reconstructed in the photographs and to select the faces that connect these corners and form a 3d model of the correct topology (figure 1). This basic set-up of camera positions is processed using a sophisticated mixture of algorithms in order to increase the stability and to reduce the number of required point correspondences. Besides the standard 2d-2d approach (8 points algorithm), a 2d-3d approach, and a 3d-3d registration approach are used [Hartley00, Faugeras93]. The appropriate algorithm is chosen automatically. With this approach, the calibration of the cameras is performed so that their position, direction, and lens parameters become known. Subsequently, the 3d positions of the corners are computed and the surface texture is extracted from the photographs automatically. By including example images from each TV camera in this process, the positions and lens parameters of the TV cameras are estimated as well, while additional images from the same camera position can be included fully automatically.

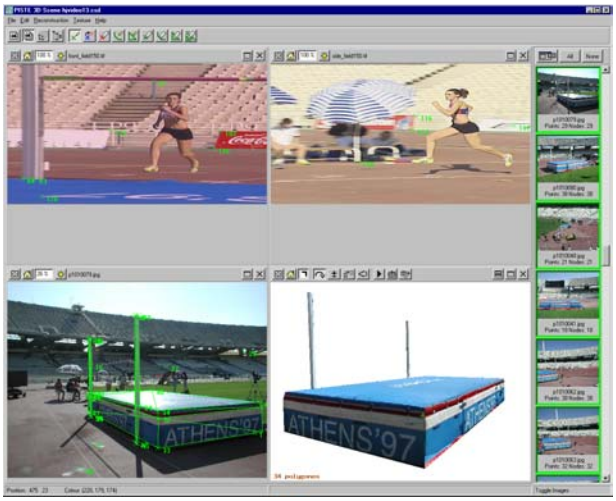

Figure 1. 3d reconstruction from multiple images.

However, video footage has to be processed in the calibration step in order to include the broadcasting cameras in the reconstruction process. By assuming a constant camera position for the broadcasting cameras, the first-order primitive on which the calibration is based is not a single image, but rather an already stitched panorama consisting of a video sequence [Coorg98]. 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 [Triggs00]. 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.

4. Re-Calibration of Broadcasting Cameras

Whenever a camera is moved to a different orientation in 3d space, a re-calibration is required in order to establish the relationship to the coordinate system of the 3d reconstruction. Fortunately, in case of static camera positions only pan, tilt, and zoom varies, so that a simplified approach can be pursued here. Using a pre-calibrated panorama image stitched from video footage from the same camera position, we only compute the orientation of the camera within this panorama (figure 2).

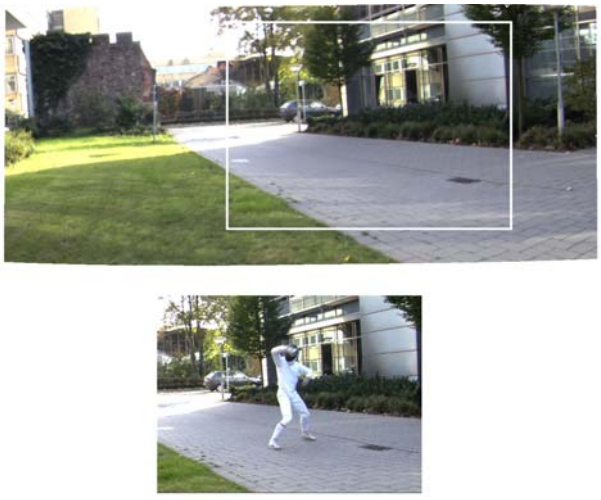

Figure 2. Registration of single video fields with pre-calibrated environment map.

From the known position of each video frame within the panorama, the camera calibration in 3d space can be derived trivially. Besides the low computational cost, this approach has the advantage of not requiring other images from different viewing directions, i.e. the background shown in the video footage from different cameras does not need to overlap. Moreover, each single video frame can be processed independently, so that parallelisation can be used in order to speed up the process.

5. 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.

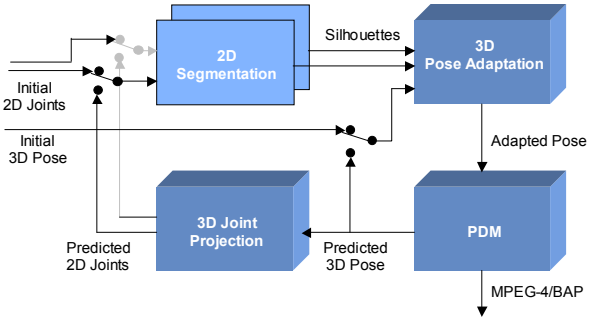

Figure 3. 3d pose estimation overview.

The images of each sequence are processed by a chain which incorporates a number of computer vision techniques (see figure 3). At first, the athlete's silhouette is determined in each view by a seed region growing algorithm [Sifakis01, Adams94]. 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 joints of the adapted body model MPEG-4 body animation parameter (BAP) are derived and used to animate an avatar at the receiver. In order to perform this step 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.

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

5.1 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, Heap96] 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.

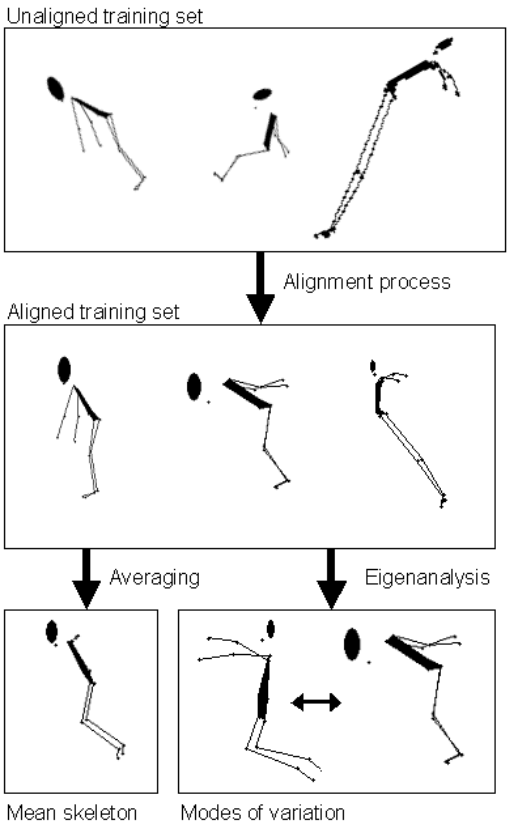

Figure 4. Creation process of a PDM.

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).

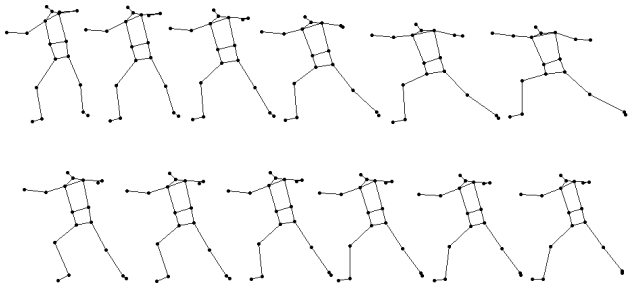

Figure 5. 1st and 2nd mode of variation for epee fencing.

Sorting the eigenvalues by its numerical quantity, it is possible to represent any skeleton as a linear combination of all eigenvectors. 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. The PDM can also be used to perform this 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.

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. Predicted skeletons can be validated, if they are within given deformation limits, and corrected to the nearest possible pose, if they are recognized as invalid ones. This predicted 3d skeleton than is projected to the image planes and used as a seed in the segmentation module.

5.2 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.

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.

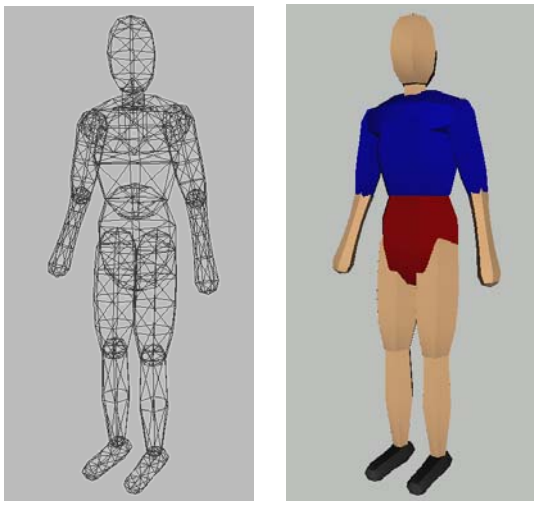

Figure 6. Wireframe and shaded representation of 3d body model.

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.

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 7 depicts an overlay of observed and synthetic silhouettes in two views. Here, green areas indicate parts of the observed silhouette which are not covered by the synthetic ones, red areas indicate the vice versa situation. The label area outside the silhouette (green areas) and the uncovered silhouette area (red areas) in each view indicate an erroneous pose (figure 7b). The pose adaptation is carried out by varying position and orientation of each body part in order to minimize sum of these areas over all available views by Powell's minimisation strategy [Press 92].

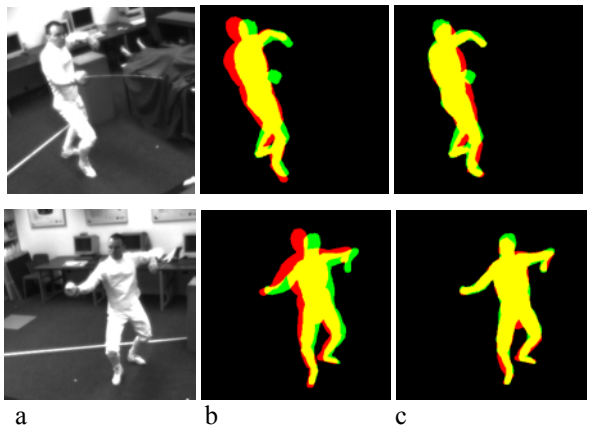

Figure 7. Input images, differences between synthetic and observed silhouettes before and after pose adaptation.

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. Subsequently, the other body limbs are optimised. The optimisation order is indicated by the numbers attached to figure 8. Final adaptation results for the field shown in figure 7a can be seen in figure 7c.

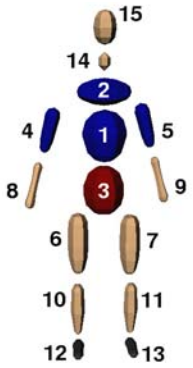

Figure 8. Input images, differences between synthetic and observed silhouettes before and after pose adaptation.

6. Business Aspects

The generated 3d sequences will be typically created by specialised service providers which buy the required

video footage externally. As no additional devices are required on the TV cameras used to acquire the video footage, it is not necessary to use dedicated cameras on site. Besides serving the future markets for 3d content in MPEG-4-based television, the approach described in this paper can be utilised in other important contexts:

- Biometric analysis of athletes in the championship situation for training purposes. Existing motion capturing systems require markers on the athlete's body and a controlled environment, which is contradictory to the championship situation, so that the most extraordinary performance defies analysis.
- The reconstructed sequences can be rendered in high quality and shown as additional analysis in conventional sports TV.

The algorithms described in this paper also offer the opportunity of product differentiation in the resulting software itself. As the Point Distribution Model requires appropriate training data for each kind of sports, the corresponding data sets will result in sports-specific complementary packages. Similarly, the different shape of the typical athlete and the additional objects involved (balls, rackets, etc) require adapted models to be used in the pose adaptation step.

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. 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.

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.

[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.

[Demiris01] 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.

[Faugeras93] O. Faugeras. Three Dimensional Computer Vision - a Geometric Viewpoint, AI MIT Press, Cambridge MA, 1993.

[Hartley00] 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

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

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

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

[Triggs00] 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

[Walczak02] K. Walczak. X-VRML – XML Based Modeling of Virtual Reality. SAINT-2002 Proceedings, 2002.