REAL-TIME 3D RECONSTRUCTION AND POSE ESTIMATION FOR HUMAN MOTION ANALYSIS

Holger Graf[††], *Sang Min Yoon*[†], *Cornelius Malerczyk*[††],

[†] Graphisch Interaktive Systeme, TU-Darmstadt (GRIS)
[††]Fraunhofer Institut für Graphische Datenverarbeitung (IGD)
Fraunhoferstraße 5, D-64283 Darmstadt, Germany

ABSTRACT

In this paper, we present a markerless 3D motion capture system based on a volume reconstruction technique of non rigid bodies. It depicts a new approach for pose estimation in order to fit an articulated body model into the captured real-time information. We aim at analyzing athlete's movements in real-time within a 3D interactive graphics system. The paper addresses recent trends in vision based analysis and its fusion with 3D interactive computer graphics. Hence, the proposed system presents new methods for the 3D reconstruction of human body parts from calibrated multiple cameras based on voxel carving techniques and a 3D pose estimation methodology using Pseudo-Zernike Moments applied to an articulated human body model. Several algorithms have been designed for the deployment within a GPGPU environment allowing us to calculate several principle process steps from segmentation and reconstruction to volume optimization in real-time.

Index Terms— Video based analysis, 3D reconstruction, pose estimation, markerless motion capturing

1. INTRODUCTION

The analysis of human motion has become an investigative and diagnostic tool for different areas such as medicine, sports or video surveillance. Ranging from human activity recognition to an in-depth analysis of motion in order to better understand normal and pathological movements, different methods have been introduced for motion analysis such as kinematic and kinetic modeling or the realization of complex capturing systems based on multiple video sensors recording the positions of markers attached to the human body. Nevertheless, the analysis of human body motion and movements within artificial generated environments still imposes major challenges to dedicated solutions. Despite the potential of marker based motion capturing systems, major hurdles for a broad acceptance have been the high costs for their installations, the requested controlled environments and the complexity of pre-processing phases for its use. At the same time, marker-free MoCap systems imply a paradigm shift away from pure marker based techniques. Whereas marker based systems provide 3D positions of markers attached to the target object, which can be captured in real-time using infrared cameras, marker-free systems rely on the deployment of an articulated human body model. This articulated body model provides "a priori" 3D positions of body segments enabling a proper association of poses as well the identification of individual body segments which allows for the extraction of kinematic information. A major bottleneck is the real-time processing of the underlying video analysis, segmentation and reconstruction steps in order to facilitate fast model fitting and analysis cycles.

Thus, in this paper, we present a new approach for a real time 3D markerless motion capture system applied to the analysis of three dimensional human body movements. It addresses typical motion analysis challenges within sports science and evaluates on the joints' angle ratios. Based on a multiple camera set-up, our workflow starts from an adequate segmentation of video streams, a robust background and silhouette extraction enabling a subsequent 3D real time reconstruction of the target objects' volumes and a new approach to accurately estimate the 3D pose of a human body model based on a statistical elevation using Pseudo-Zernicke moments

2. RELATED WORK

Edward Muybridge pioneered the work on motion capturing with his famous experiments in 1887 called "Animal Locomotion" [1]. Since then, a lot of research changed methodologies and techniques of motion capturing and its analysis. With recent technology developments in the area of hardware accelerated computer and cameras, marker based motion capture systems, e.g. [2][3], provided an accurate position of target objects. In order to overcome physical limitations such as installation restrictions and imposed constraints for user's activity capturing, maker-free motion capture systems, e.g. [4][5] have been introduced to overcome those drawbacks. Our approach for markerless motion capturing and its reconstruction for 3D interactive analysis, comprises different techniques addressing real-

time photo-realistic 3D reconstruction and model based 3D pose estimation. 3D reconstruction research started early on by stereo vision based techniques proposed by [6] being extended into a multiple camera environment, e.g. [7][8]. Those methods are designed to reconstruct depth maps from particular viewpoints though being not suitable for a full 3D scene reconstruction. Image based visual hull reconstruction (IBVH) proposed by [9] is a real-time 3D scene reconstruction technique from multiple view images. The algorithm does not solve a correspondence problem. Instead, it calculates the convex hull of silhouettes within all view images. A voxel coloring method [10] did resolve the problem of reconstructing concave objects.

In view of best-for-fit pose estimation of an articulated model onto a real body pose, a large number of papers have been published, e.g. [11][12] to name few of many. Typically, model based 3D human pose estimation methods are separated into two approaches: appearance-based methods [13] and part-based methods [14]. Those classes differentiate in either using the full human appearance information or exploiting parts of a human body such as face, torso, and limbs for a model fitting.

3. VISION BASED 3D RECONSTRUCTION OF NON RIGID BODIES

3.1 Kernel Density Estimation based Background Subtraction

Our methodology relies on a high quality video segmentation of target objects. We use a background subtraction technique to detect the deformable objects in the scene by comparing each new frame to a pre captured model of the scene background. Here, we apply a non-parametric technique for background modeling and foreground extraction. Our approach is based on a kernel density estimation applied to the probability density function of the intensity of each pixel within each image. Kernel density estimation based background modeling aims at capturing and storing recent information about the image sequence, continuously updating this information in order to capture fast changes in the scene background [15].

The intensity distribution of a pixel can change quickly. So we can estimate the density function of this distribution at any moment of time given only very recent history information if we want to obtain a sensitive detection. Using the recent pixel information, the probability density function of each pixel depending on intensity value $I_t(x,y)$ at time t and can be non-parametrically estimated using the kernel, K as

$$pdf(I_t) = \frac{1}{N}\sum_{i=1}^{N} K\left(I_t - I_i\right)$$

where N is the recent pixel information used to compare the current image's pixel information. We choose our kernel estimation function to be a Gaussian kernel for color images. Figure 1 shows the background subtraction from multiple color and gray images (source: HumanEva[1] dataset).

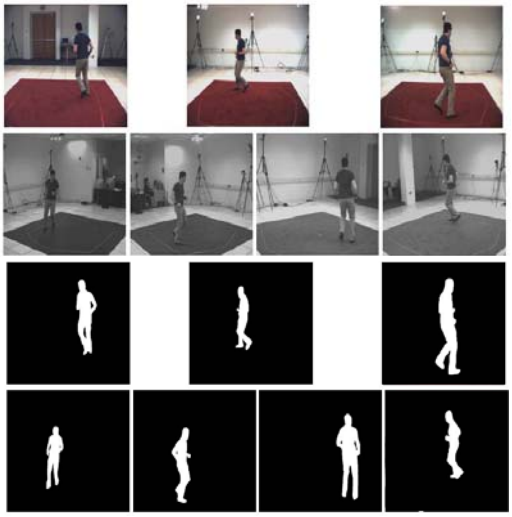

Figure 1 Results of a silhouette extraction using our kernel density estimation based background subtraction (basis for the appearance model and used for several downstream processes; color plates)

3.2 3D Reconstruction of Target Objects

Our 3D reconstruction methodology is based on a multiple video stream captured from different calibrated camera positions. We exploit the presented foreground information as extracted appearance model of the target object.

Based on this input information we reconstruct the external surface of the volume based on the IBVH being improved by the voxel coloring technique as presented by [10]. The methodology we present in this paper combines both approaches with an additional tracking component. Thus, we use the first technique in order to reconstruct a very coarse 3D shape based on only few images, as it is error prone to the quality of the images, position of each viewpoint, quality of camera calibration and the complexity of the object's shape. Afterwards we deploy the voxel coloring which reconstructs the radiance or color at the surface points by projecting every voxel to each image plane. Our proposed 3D reconstruction methodology continuously tracks the 3D boundary of target object and carves the voxels by checking the color consistency within the captured and tracked 3D boundary. This leads to an efficient and accurate method improving previous approaches deployed within a large environment. Practicability aspects and a high degree of parallelism of the used techniques allow the mapping of several sub-processes and computing steps to a parallel computing architecture and hence have been implemented on CUDA. Figure 2 shows the concept of the configuration of a 3D lattice by tracking the target object and its inverse projection within the 3D scene using n calibrated static cameras. Here, we

[1] http://vision.cs.brown.edu/humaneva/

continuously track the center of gravity $g_1,g_2,...,g_n$ of our appearance model in each image and calculate the G points in the 3D scene which we get by the intersection of n 3D rays. We extract the 3D lattice by combining the silhouette images of the target object using the camera calibration information in order to set the visual rays within the 3D space for all silhouette points, which define a generalized cone. The 3D lattice in a whole scene is then determined by its intersection of those cones. Using the 3D lattice, we deploy the photo-consistency measure to determine if a certain voxel X does or does not belong to the object being reconstructed.

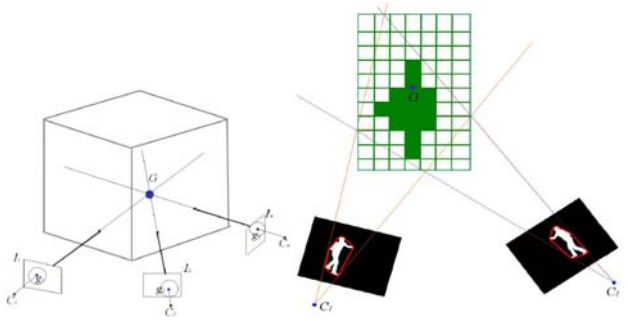

Figure 2 3D lattice configuration by tracking the 3D boundary of a target object (color plates)

The following figure shows the results obtained in different steps.

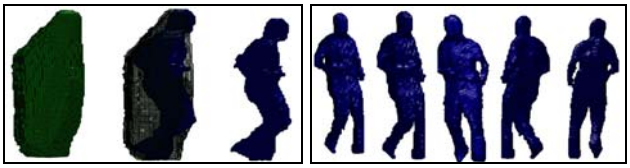

Figure 3 Results (color plates): multiple input images ref figure 1; 3D lattice and 3D reconstructed object within 3D lattice (bottom left); body model different viewpoints (bottom right)

4. 3D POSE ESTIMATION

In order to comply with the paradigm change fitting an articulated model into the information provided by multiple video streams, we exploit a 3D pose estimation technique enabling an efficient mapping of body segments and joints into the reconstructed volume. Within our appearance model, statistical moments are used as segment descriptors. The approach is based on the Pseudo-Zernicke Moments (PZM), which are used as segment descriptors. Those moments are typically the statistical expectation of certain power functions of a random variable. In general, moments of order $(p+q)$ for a continuous function $f : R^2 \to R$ can be written as

$$M_{pq} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \phi_{pq}(x,y)\, f(x,y)\, dxdy$$

with $p,q=0,1,2,...$ and $\phi_{pq}(x,y)$ the basis function.

Adapting this to image processing tasks, $f(x,y)$ may be understood as pixel intensities of the segmented binary image P_{xy}. The discretization leads to

$$M_{pq} = \sum_x \sum_y \phi_{pq}(x,y) P_{xy} \quad , p,q=0,1,2,...$$

Teh and Chin [16] evaluated various different moments and as a result they could show that Zernike moments and especially Pseudo-Zernike moments outperform other moments in terms of overall performance. Pseudo-Zernike moments are an adaptation of standard Zernike moments [17],[18] with orthogonal radial polynomials as basis functions of order p and repetition q

$$W_{pq}(r\cdot\cos\theta, r\cdot\sin\theta) = R_{pq}\cdot e^{iq\theta}$$

with $p \geq 0, q \leq p, q \in Z$. Radial polynomials are calculated by

$$R_{pq} = \sum_{s=0}^{p-|q|} (-1)^s \frac{(2p+1-s)!}{s!(p-|q|-s)!(p+|q|+1-s)!}\cdot r^{p-s}$$

with $\langle W_{nm}, W_{pq}\rangle = \frac{\pi}{n+1}\cdot\delta_{np}\delta_{mq}$ Pseudo-Zernike moments are defined as

$$P_{pq} = \frac{n+1}{\pi}\cdot\langle f, W_{pq}\rangle = \frac{n+1}{\pi}\iint f(x,y) W_{pq}^{+}(x,y)\, dxdy$$

The reconstruction of the body movements is based on the minimization of the difference between artificial and real silhouettes within all camera images. Three basic constraints are used to achieve real time capability:

1. a minimal setup of the human skeleton to achieve a reduction of the dimension of the search space.
2. a discretization of possible movements into a predefined set of individual body poses.
3. time consuming generation of artificial silhouettes is done prior to the online reconstruction phase.

A typical minimal setup for both rendering and reconstruction tasks uses only 15 joints with overall 32 degrees of freedom to describe the human skeleton: Root joint (for the position and orientation of the human body, sacroiliac (separate lower and upper body part), hips, knees and ankles, shoulders, elbows and wrists and the skullbase. This minimal kinematic model of a human body used as well for the generation of artificial silhouettes. The algorithm thus evaluates as follows: In a first step a database is generated after each new camera calibration consisting of vectors for each predefined pose, where each vector has:

- all Pseudo-Zernike moments of all four artificial silhouettes,
- used angles for the root node of the avatar,
- all internal joint rotation of the actual pose and
- a reference to silhouette images, which are saved during building the database.

Each of these vectors is one record of the database, which is used during the online recognition process to find the best matching pose for a set of silhouettes derived from real camera image data. It is reasonable, that the best match will be found, if the Euclidean distance between two vectors is minimal. A fast approximation to find the best matching vector is to sort the database by the first Pseudo-Zernike moment. Here, the database is sorted by the quantity of the first moment. During the recognition phase, the vector with the closest first moment is located and for a given neighborhood of n vectors (e.g. 1000) the Euclidean distances between the vector of real silhouettes and database vectors are calculated.

5. EXPERIMENTS & CONCLUSION

We setup our proposed methodology using a Pentium 4 1.2 GHz CPU and equipped with recent NVIDIA Geforce 8200 graphics board (CUDA support). The multiple camera system is using four color Imaging Source cameras at 640x480 resolution for each captured image stream. We lead some experiments to analyze the athletes motion in real-time. The capturing square space is 2m x 12m in order to account for an athlete's sprint start. In a first experimental set-up, we also validated our results at a museum exhibition for innovative computer applications in sports science. Figure 4 shows the set up and illustrates the different steps, i.e. input images, 3D reconstruction and 3D pose estimation.

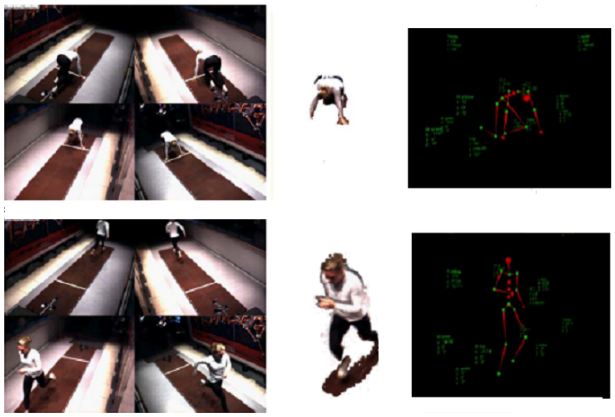

Figure 4 (color plates) Snapshot of the experimental set up (left); results of 3D reconstruction (middle), 3D pose estimation (right)

The benchmark during the installation of the exhibit in order to evaluate for the real-time performance (ref. table 1):

Category	module	running time (ms)
3D Reconstruction		189
	Segmentation	38
	Tracking	15
	3D reconstruction	136
3D Pose Estimation		74
	Segmentation	38
	Pseudo-Zernike pose estimate	36

Table 1 Timing benchmarks needed for sub processes

This paper presents a novel framework for markerless motion capturing based on an optimized 3D reconstruction and 3D pose fitting technique into a multiple video stream. It introduces a new tracking mechanism for an improved image based volume reconstruction and an appearance model for pose estimation based on PZM. Some bottlenecks have been observed so far in our approach: The quality of 3D reconstruction and 3D pose estimation depends significantly on the noise within the segmented images. A pretty good "sharp" segmentation of the target object could lead to the distortion of the 3D volume reconstruction (especially due to light foreground and bright background). In turn, the 3D pose estimation leads to a significant higher pose detection facilitating smooth transitions of the skeleton animation (figure 4 above). A less sharp segmentation leads to a good reconstructed volume but a lower quality in pose fitting (figure 4 below).

6. REFERENCES

[1] Haas, R. Bartlett, "Muybridge: Man in Motion". *University of California Press*, 1976

[2] MoCap-Systems - Motion Analysis, Vicon, Simi: Marker based tracking systems. www.motionanalysis.com, www.vicon.com, www.simi.com/, June, 2005

[3] T. Moeslund, and E. Granum, "A survey of computer vision based human motion capture". *Vision and Image Understanding*, 81(3):231-268, 2001

[4] P. Kaimakis, and J. Lasenby, "Markerless motion capture with single and multiple cameras". *Proceeding of ICIP*, 2004

[5] S. M. Yoon, C. Malerczyk, and H. Graf, "3D skeleton extraction from volume data based on normalized gradient vector flow". WSCG, 2009

[6] D.C. Marr, and T. A. Poggio, "A computation theory of human stereo vision". *Proceeding of the Royal Society of London*,pp.301–328, 1979

[7] M. Okutomi, and T. Kanade, "A multiple-baseline stereo". *IEEE Transactions on PAMI*, (15): 353–363, 1993

[8] S. B. Kang, R. Szeliski, and J. Chai, "Handling occlusions in dense multi-view stereo". *Proceedings of Computer Vision and Pattern Recognition*, pages 103–110, 2001

[9] W. Matusik, C. Buehler, S. J. Gortler, and L. McMillan, "Image-based Visual Hulls". *Proceedings of ACM SIGGRAPH*, 2000

[10] S. M. Seitz, and C. M. Dyer, "Photorealistic scene reconstruction by voxel carving". *Proceeding of Computer Vision and Pattern Recognition*, 1997

[11] T. B. Moeslund, A. Hilton, and V. Krueger, "A survey of advances in vision-based human motion capture and analysis". *Computer Vision and Image Understanding*, 99–126, 2006

[12] I. Kakadiaries, and D. Metaxas, "Model-based estimation of 3-D human motion".*IEEE Transactions on PAMI*,22(12):1453–1459, 2000

[13] C. Papageorgiou, T. Evgeniou, and T. Poggio, "A trainable pedestrian detection". *IEEE Intelligent Vehicles Symposium*, 1998

[14] P. F. Felzenswalb, and D. P. Huttenlocher, "Efficient matching of pictorial structures". *Proceedings of CVPR*, 2000

[15] B. Han, D. Comaniciu, and L. Davis, "Sequential kernel density approximation through mode propagation". *Proceedings of ECCV*, 2004

[16] C. H. Teh, and R. T. Chin, "On image analysis by the methods of moments". *IEEE Transactions on PAMI*, pp. 496–51, 1998

[17] F. Zernike, "Beugungstheorie des Schneidenverfahrens und seiner verbesserten Form, der Phasenkontrastmethode". *Physics 1*, pages 689–701, 1934

[18] A. Khotanzad, and J.J.-H Liou., "Recognition and pose estimation of unoccluded three-dimensional objects from a two-dimensional perspective view by banks of neural networks". *IEEE Transactions on Neural Networks*, 7(4): 897 – 906, 1996