ANALYZING HUMAN BODY 3-D MOTION OF GOLF SWING FROM SINGLE-CAMERA VIDEO SEQUENCES

Ibrahim Karliga and Jenq-Neng Hwang

Department of Electrical Engineering, Box 352500
University of Washington, Seattle, WA 98195

ABSTRACT

We present an algorithm for analyzing the human body 3-D motion of golf swing from single-camera video sequences. As the first step, the human body used for the analysis is automatically extracted using a video object segmentation technique. Once human body is extracted, this two-dimensional information is utilized to obtain a three-dimensional body model consisting of head, upper arms, lower arms, body trunk, upper legs, lower legs and feet using an iterative 3-D fitting algorithm and Dynamic Bayesian Networks. The ultimate objective of the system is to obtain the 3-D motion information for golf swinging and comparing this information for different players regardless of the great variability caused by different camera viewing perspectives. 3-D body motion during the golf swing is computed for all segments of the body. This system will allow the spatial-temporal relationship of each body segment, as they make their transition, be thoroughly studied, and enable the parameters for different players to be compared, as well.

1. INTRODUCTION

The recent technological advances in real-time capturing, transferring and processing of images/video on standard hardware systems has established the visual analysis of human movement in video sequences as one of the main application domains of computer vision. In this respect, sports video analysis has been one of the emerging topics as it has great commercial potentials, as well as the immediate need to store, index and filter the sports video data accumulated to facilitate the search for specific contents. Some of the major research issues in sports video analysis have been sports tactics summarization, ball/player tracking, game highlight extraction, computer-assisted refereeing, content insertion, personalized training, etc. In these literatures, there have been primarily two approaches in sports video content analysis. The major one is the frame-based approach where features are extracted directly from frames, and the other one is the object-based analysis to interpret the object behavior in a very fine level [1-5].

A variety of sports including baseball, soccer, football, and tennis [6-9] have been used to demonstrate new ideas in many sports video content analysis projects. The methods employed include statistical methods, rule based methods, Hidden Markov Models, and Dynamic Bayesian Networks.

The solution of the problem of 2-D to 3-D inference has intrinsic difficulties due to the loss of 3-D information in projection to 2-D images. Recovering 3-D human model from 2-D video sequences gets even more difficult due the variability of human appearance, the high dimensionality of articulated body models, self occlusion, and the complexity of human motion. In [10], the 3-D body tracking is treated as an inference problem where the inherent 3-D ambiguity of 2-D video was solved by relying on prior knowledge about human motion learned from training data to reconstruct the 3-D motion of human subjects from single-camera video. In [11], the 3-D motions of articulated models are computed from 2-D correspondences using an iterative batch algorithm. It estimates the maximum *a-posteriori* trajectory based on 2-D measurements subject to a number of constraints: kinematic constraints based on a 3-D kinematics model, joint angle limits, dynamic smoothing and 3-D key frames specified the user. In [12], the 2-D projection of a 3-D arm model was compared against 2-D image features for updating the EKF filter for 3-D tracking. A 3-D human body model, skeleton extraction and iterative closest point algorithms (ICP) were utilized to estimate the 3-D information from a monocular video sequence in [13]. In this paper, we propose an innovative video analysis system for analyzing 3-D motion of golf swing based on a coarse-to-fine grain video object analysis framework similar to [14]. The proposed system also employs a pattern analysis methodology by modeling the video object behavior using Dynamic Bayesian Networks. This system shows great promise in recovering 3-D human motion from 2-D video sequences regardless of variability of human appearance, camera perspective, and the high dimensionality of articulated body models, self occlusion, and the complexity of human motion.

The remainder of the paper is organized as follows. In Section 2, the system framework including the video object segmentation, the extraction of the lower arms of the players, 3-D Human Body Model, the fitting of the human

body to 3-D Model, as well as the Dynamic Bayesian Networks Model is explained. Section 3 includes experimental results followed by conclusion in Section 4.

2. GOLF SWING HUMAN BODY MOTION ANALYSIS SYSTEM

2.1. System Framework

As shown in Figure 1 that our automatic algorithm for analyzing the 3-D motion of golf swing starts with a video object segmentation technique, extracting of human body, a 3-D body model consisting of head, upper arms, lower arms, body trunk, upper legs, lower legs and feet using an iterative 3-D fitting algorithm and Dynamic Bayesian Networks.

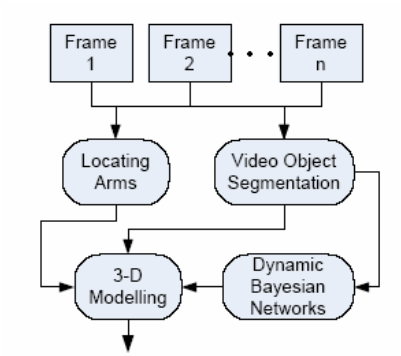

Fig.1: The flow chart of the proposed system.

2.2. Human Body (Video Object) Segmentation

The first step of the overall 3-D fitting algorithm is segmenting out the human body from each frame. Initially, an algorithm based on mean shift segmentation [15] is used to perform region merging. Then, using multiple frames, the regions with significant motion are labeled as belonging to the video object, also making the assumption that the player is located roughly in the center of the image frame. Fig. 2 shows the performance of this segmentation method.

Fig.2: Segmentation Performance for Different Frames.

2.3. Extracting Lower Arms of the Players

It's clear that the video object segmentation should be accurate enough in order to get a reliable 3-D Model. Nevertheless, since only contour information is utilized to fit the 3-D model for the initial video object, the fitting performance degrades considerably when self-occlusion by the arms occurs. Hence, in these situations, extra information has to be provided to the fitting algorithm to enhance the performance.

To address this issue, we locate the arms in the first segmented video object by taking advantage of the information that in the beginning of golf swing the arms should be roughly V shaped (Fig. 3). First, region merging using color information is performed, and for each merged region the feature vectors that include the major axis length, the minor axis length, the orientation, and the centroid is evaluated. Finally, the locations of the lower arms are determined using this feature vector.

Fig.3: Lower arms extracted.

2.4. 3-D Human Body Model

As 3-D body model, we use the one similar to the one used in [14]. It consists of head, upper arms, lower arms, body trunk, upper legs, lower legs and feet where they are modeled by geometrical objects of appropriate sizes and dimensionality ratios, and the joints, (the neck, shoulders, elbows, thigh sockets, knees and ankles) are modeled by spheres with appropriate sizes (Fig. 4). The joints do not participate in the motion, and they are used only to fill the gaps between body parts to allow no holes or gaps in the 2-D shapes after the 3-D to 2-D projection. The body parts are sized according to appropriate ratios to the length of the

body trunk and the ratios are adjustable in certain ranges to fit body shapes of different players.

The motion of the human body is a typical articulated object motion where the motion of each body part can be transformed to the global coordinate system by concatenation of the local coordinate systems. The system has a total of 22 degrees of freedom. 6 of them are for the whole body, 2 of them for the head, 2 for the upper arms and thighs, 1 for the lower arms and lower legs, and 1 for the feet. The system has a total of 58 independent parameters including the body part size ratios and 3-D model coordinates.

Fig.4. 3-D human body model used with different configurations.

2.5. Fitting Initial Video Object (Human Body) to 3-D Model

As soon as the video object representing the player is extracted and the location of the lower arms are determined, the next step is fitting this to the 3-D model described above using an iterative Nelder-Mead's method [14]. Here, the camera is assumed to be located vertically to the image plane, and there is no rotation between the world coordinate system and the camera coordinate system. Also, the *a priori* information obtained about the orientations of the two lower arms is imposed as additional constraints and fed directly into the 3-D model before starting iterative merging algorithm. This 3-D model with imposed constraints comprises the initial version of the 3-D model.

The 3-D model is projected from camera model to the image plane using the following transformation [14]:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} + \frac{f}{z_c} \begin{pmatrix} x_c \\ y_c \end{pmatrix} \quad (1)$$

and, the silhouette of the projected 3D model and that of the video object are compared using the "*exclusive OR"* operation. The matching error is represented in terms of number of non-matching pixels, and the optimization algorithm tries to converge to the global minimum by changing the appropriate variables in the 3-D model in a coarse to fine fashion using Nelder-Mead's method [14].

2.6. Fitting Rest of VO to 3-D Human Body Model

Once the 3-D model for the human body in the first frame is obtained, this is used as the initial 3-D model for the video object in the next frame. The same fitting method, i.e., iterative Nelder-Mead's method, and the matching criteria are employed to find the final model. This procedure goes on recursively until the very last frame of the video sequence. Also, as an additional channel of information to the fitting algorithm, we employ a pattern analysis methodology by modeling the video object behavior using Dynamic Bayesian Networks (DBNs), details of which are explained below [16]. This methodology requires a batch learning process in advance, and the resulting information for the location of the arms, and legs for each frame is fed directly into the 3-D fitting algorithm as an additional constraint.

The initial process for DBN modeling is feature extraction. We use geometrical features extracted from the shape of VOs. For each VO, we extract the temporal position change of the head, hands and legs following the algorithm based on skeletonization of human body shape. It starts with the binarization of the VO. Next, the skeleton of the binarized shape of VO is extracted based on the morphological operation. The center of gravity (COG) of the shape is identified, and an axes system with the COG as the origin, the VO region is divided into four quadrants (I, II, III, and IV). Finally, the set of ''end points'' of the skeleton are detected and normalized to offset the effect of different body sizes. The mean x and y coordinates of the normalized end points in each quadrant are obtained and used as the features [16].

Having obtained the feature vectors, DBN modeling is initiated. There are five hidden nodes and four observation nodes in each time slice. The hidden nodes represent the states of head, hands and feet in each time slice. Each hidden node has four states representing their position relative to COG, upper-right (quadrant I), upper-left (quadrant II), lower-left (quadrant III), and lower-right (quadrant IV). Corresponding to the extracted features, there are four observation nodes representing the mean x and y coordinates in four quadrants.

During the learning stage, the value of hidden nodes (i.e., in which quadrant the limbs and head are located) are obtained through manual annotation for all the training video sequences. The values of observation nodes, i.e., feature vectors in each quadrant were then calculated, and the values of both hidden nodes and observation nodes were fed into the DBNs. After the parameters of the specific DBNs were learned, the feature vectors of the testing sequences are fed into these DBNs to compute the log likelihood for classification and further perform decoding to obtain the maximum probable explanation (MPE). By calculating the MPE of a DBN, each end point can be successfully grouped into one of the body part (head, left/right hands, left/right legs).

3. SIMULATION RESULTS

A total of 38 golf swing sequences were used to learn the DBN and test the algorithm. Due to the limited space, we present the results for only one sequence. Figure 4 shows

the skeletons of the segmented video objects marked with the end points in Figure 5.

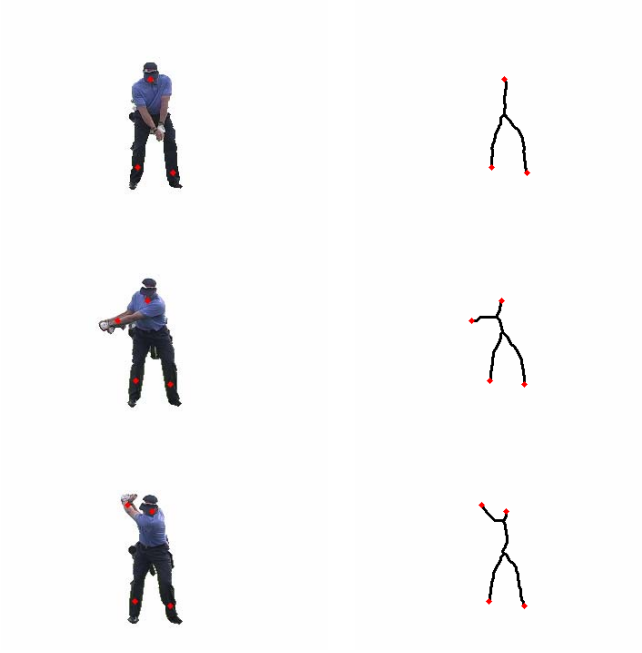

Fig. 5: Skeletonization Performance for Different Frames

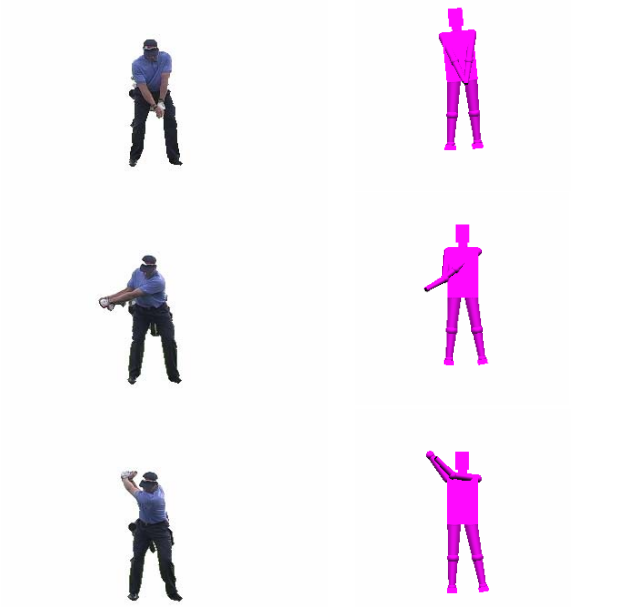

Fig.6: 3-D Model Fitting Performance for Frames in Fig. 5

4. CONCLUSION

We propose an automatic 3-D motion of golf swing analysis system which effectively integrates a video object segmentation technique based on mean shift region tracking, the locating of human body parts based on DBN modeling, and a 2-D to 3-D body fitting optimization algorithm. We note that, the algorithm especially differs from the other 3-D sports analysis algorithms as it utilizes a pattern analysis approach by modeling the video object behavior using Dynamic Bayesian Networks. The system has shown to be capable of grasping the 3-D human body structure in golf swinging. This system will allow the spatial-temporal relationship of each body segment, as they make their transition, be thoroughly studied, and enable the parameters for different players to be compared, as well.

REFERENCES

[1] T. Lin and H.-J. Zhang, Automatic video scene extraction by shot grouping, Proc. 15th ICPR, Vol. 4, Sept. 2000, pp. 39-42.

[2] W. Tavanapong, Z. Junyu, Shot clustering techniques for story browsing, Multimedia, IEEE Trans. on, pp. 517- 527, Vol. 6-4, Aug. 2004

[3] E. Veneau, R. Ronfard, P. Bouthemy, "From video shot clustering to sequence segmentation," in Proc. 15th Int. Conf. Pattern Recognition Barcelona, Spain, vol. 4, Sept. 2000,

[4] H. Sundaram and S. F. Chang, "Video scene segmentation using audio and video features," in Proc. IEEE ICME 2000

[5] M Petkovic, V. Mihajlovic, W. Jonker, Techniques for automatic video content derivation, Proc. ICIP 2003. Vol. 2, 14-17 Sept. 2003,

[6] Chang, P., Han, M., Gong, Y.: Extract highlights from baseball game video with hidden Markov models, Proceedings of the International Conference on Image Processing (ICIP '02). (2002)",

[7] J. Assfalg , M. Bertini , C. Colombo , A. D. Bimbo, Semantic Annotation of Sports Videos, IEEE Multimedia, v.9 n.2, p.52-60, April 2002.

[8] N. Babaguchi, Y. Kawai, Ogura, T.; T. Kitahashi, Personalized abstraction of broadcasted American football video by highlight selection, Multimedia, IEEE Trans. on, Volume 6, Issue 4, Aug. 2004 Page(s):575 - 586

[9] G. Sudhir, John Chung-Mong Lee, Anil K. Jain: Automatic Classification of Tennis Video for High-Level Content-Based Retrieval. CAIVD 1998: 81-90

[10] N. Howe, M. Leventon, W. Freeman, Bayesian Reconstruction of 3-D Human Motion from Single-Camera Video, NIPS'99, Denver, USA, Nov.1999.

[11] D. E. DiFranco, C. Tat-Jen, J.M. Rehg, Reconstruction of 3D figure motion from 2D, Computer Vision and Pattern Recognition, 2001, pp. 307-314 vol.1

[12] L. Goncalves, P. Perona, Monocular tracking of human arm in 3D, ICCV'95, Boston, USA, Jul.1995.

[13] Sappa, A., Aifanti, N., Malassiotis, S., Strintzis, M.G., Monocular 3D human body reconstruction towards depth augmentation of television sequences, ICIP'03, Barcelona, Spain,

[14] Y. Luo, J-N Hwang, A Comprehensive Coarse-To-Fine Sports Video Analysis Framework to Infer 3D Parameters of Video Objects with Application to Tennis Video Sequences, IEEE, ICASSP 2005.

[15] Y. Cheng, "Mean shift, Mode seeking, and Clustering," IEEE Tr. on Pattern Analysis and Machine Intelligence, Vol. 17, No. 8, pp. 790- 799, August 1995.

[16] Y. Luo, T. D. Wu, J-N Hwang, Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks, Computer Vision and Image Understanding 92 (2003) 196–216.