A COMPREHENSIVE COARSE-TO-FINE SPORTS VIDEO ANALYSIS FRAMEWORK TO INFER 3D PARAMETERS OF VIDEO OBJECTS WITH APPLICATION TO TENNIS VIDEO SEQUENCES

Ying Luo, Jenq-Neng Hwang
Information Processing Lab. (IPL)
Dept. of Electrical Engineering
University of Washington
Seattle, Washington 98195
{luoying, hwang}@ee.washington.edu

ABSTRACT

In this paper, we present a novel video content analysis system. An innovative 2D to 3D parameter inference algorithm is presented. It is applied to the tennis player body shape modeling, after a coarse-to-fine analysis on real world sports video sequences. As the first step, the video shots are classified in coarse level. Only shots containing appropriate body shape size are retained for the fine-level analysis. The fine-level analysis begins with a video object (VO) segmentation stage to obtain the player body shapes. The VOs then undergo training and testing stages. The training VOs are classified into serving and non-serving classes by Gaussian mixture modeling (GMM). The VOs in serving class are further clustered and the corresponding 3D parameters of a human body model are obtained manually for each cluster center. For a testing VO sequence, the VOs that contain servings are found by GMM and the initial 3D parameters are fitted to the closest matches to the cluster centers. Based on the initial guess, an innovative multidimensional optimization procedure is employed to obtain the 3D parameters. Experiments are performed on broadcasted tennis games and promising results are obtained.

1. INTRODUCTION

With the popularity of digital video recording, video content analysis has become a basic requirement for the indexing, retrieving and analysis of video sequences. There are two methodologies in video content analysis: one is through the features extracted from frame-based approach to obtain the coarse-level knowledge of the video contents. The other is through the object-based analysis so as to interpret the object behavior in a very fine level.

Sports video clips are of specific interest in research due to their availability and moderate complexity. Most of the sports event detection efforts were in coarse-level based on frame features, such as [9, 11, 15], to name a few. Some of the projects are more concerned with the region or object based behavior analysis [13, 1, 7]. Tennis video analysis was used to demonstrate ideas in many sports video content analysis projects. In [11], the tennis video clips were used to demonstrate the clustering method in finding out the structure of the video sequences by low-level features such as colors and their derivatives. Different modeling techniques, such as rule-based methods, HMM and DBN were compared in [15] with experimental examples for tennis video sequences.

For the fine-level object-based analysis, the video objects need to be segmented first. The focused video objects in tennis video sequences are the player body shapes. Although the topic of 2D to 3D inference is rarely put in the sports video analysis, the model-based 2D to 3D inference has been the dominant approach for the human body modeling. In [2], the 2D projection of a 3D arm model was compared against 2D image features to provide update for the EKF filter for 3D tracking. A full 3D human body model was used in [17] for similar purpose. Except for tracking, 3D human body model is also used to infer the 3D information directly from 2D images. Machine learning methodology was first used to learn the probabilistic model of the 3D motion in [3]. Then the Bayesian method was employed to infer the real 3D motion of the testing sequences based on 2D measurement and a 3D skeleton model. A full 3D human body model, skeleton extraction and iterative closest point algorithms (ICP) were combined to estimate the 3D information directly from input monocular video sequence in [16].

The contributions of this paper are in two aspects. First of all, a novel coarse-to-fine framework is proposed to deal with the real world tennis video content analysis problems. By selecting appropriate shots for the fine-level analysis through coarse-level analysis, it fills the gap between coarse level analysis and fine level analysis that is present in many of the current research work. Secondly, an innovative video object (VO) based 2D to 3D parameter inference scheme is proposed and tested to be effective in this paper. Combining both VO segmentation and 3D models, the proposed method is appropriate for VOs with articulated structure and complicated motion patterns. The tennis players are used as examples and the results shown are promising.

In Section 2, we will give the detailed description of the coarse-to-fine framework and VO-based 2D to 3D inference algorithm. The experimental results are provided in the Section 3, followed by the Conclusion and Discussion in Section 4.

2. A COARSE-TO-FINE TENNIS VIDEO ANALYSIS SYSTEM

2.1. System framework

The system framework is shown in Fig 1. The system is divided into two big modules: the coarse-level module and fine-level module. In the coarse-level module, after the shot segmentation, the shots are classified into two different classes: one is the shot with large enough player body shape for further fine-level analysis, the other is all other shots. Again, the tool employed is

HMM. In the fine-level module, only the shots containing player body shapes are further processed. The body shapes are segmented out by our video object segmentation algorithms. The poses of the body shapes are further analyzed: they are first classified into serving and non-serving classes. Then the VOs in serving class are first clustered into clusters. A 2D to 3D inference is employed to infer the 3D gesture of the athlete during the serving.

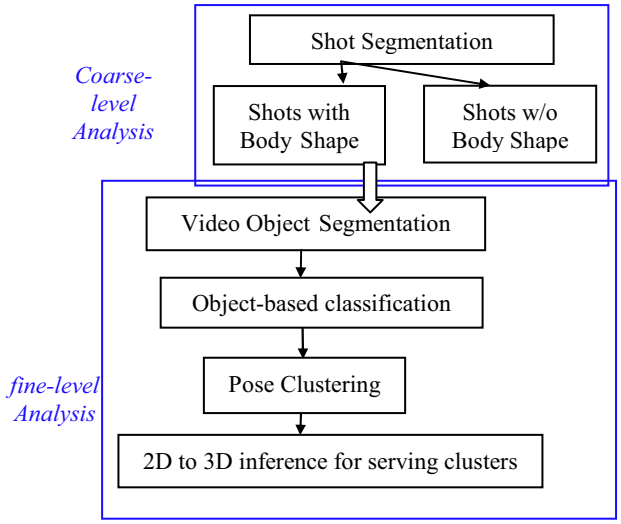

Fig. 1. The coarse-to-fine video content analysis system framework

2.2 Coarse-level analysis

The coarse-level analysis includes the shot segmentation and shot classification based on frame features.

In the previous sub-section we have mentioned that we only want to choose the shots with body shapes of appropriate size so we can analyze the shape in the fine-level analysis. We generally have four shot classes: shots with appropriate body shapes (BS), shots containing far field (FF), shots with audience close-up (AC), and shots with audience far-shooting (AF).

The AC and AF are removed first based on color and texture features. Then BS and FF shot classes are classified by Hidden Markov Models (HMMs). For details, please refer to [12].

2.3 Fine-level analysis

Only the shots containing appropriate body shapes are kept for the fine-level analysis. The fine-level analysis is based on video objects: the video objects are first segmented out and then the 2D to 3D model fitting is performed.

2.3.1 Video object segmentation

In the fine-level analysis, the first step is video object (VO) segmentation. We have continuously developed VO segmentation algorithms over the years [5, 6, 4]. The VO segmentation can be performed either fully automatic or semi-automatic based on the complexity of the shots.

2.3.2 2D to 3D inference

After obtaining the VOs, we want to infer reasonably accurate 3D pose information from the 2D video object sequence. Specifically, since we are only interested in the video objects containing the servings, there are two problems that need to be solved: to identify the serving video objects and to infer the 3D parameters of the serving poses.

2.3.3 Serving video objects classification

In the VOs, a lot of other poses are present: playing with balls, standing around, walking, running, serving, etc. To identify the serving poses, we use the Gaussian mixture modeling (GMM). The algorithm is as follows:

1. In the training stage, the parameters are obtained through training data for the Gaussian mixtures (GMs). A GM is obtained for the serving poses and another GM is obtained for all other poses. The feature used here is the classic Hu's moments of the object shape.
2. In the testing stage, maximum likelihood strategy is used for the input video object sequence. The probability of each VO to the serving pose GM and other pose GM is calculated and compared. Only when the maximum likelihood ratio of at least one VO exceeds a certain threshold, the sequence is declared to contain a serving pose.

2.3.4 2D and 3D model fitting

After the serving shots are found, we need to infer the 3D parameters from the 2D shape. Mathematically, this is a problem with multiple solutions for still images because the 3D object to 2D image region projection is a multiple-to-one projection. For the video case, however, we have better chance since the motion of the object and/or the camera can provide more 3D information about the object of interest by revealing different parts of 3D object shape through time.

2.3.4.1 The 3D human body model

As shown in Fig. 2, the 3D body model consists of: head, upper arms, lower arms, body trunk, upper legs, lower legs and feet. This body parts are modeled by geometrical objects: the head, body trunk and feet are cubes of appropriate size and dimensionality ratio. The arms and legs are cylinders with different radii at two ends. All the joints, including the neck, shoulders, elbows, thigh sockets, knees and ankles, are modeled by spheres with appropriate sizes. Unlike other body parts, these joints do not participate in the motion: they are only used to visually fill the gaps between body parts so there will be no holes or gaps in the 2D shapes after the 3D to 2D projection.

The model can be freely resized: the length of the body trunk cube is used as a major reference size parameter for all other body parts. The body parts are sized according to appropriate ratios to this parameter and the ratios are adjustable in certain ranges to fit body shapes of different players. Examples are shown in Fig. 2 (a) and (b).

The motion of the human body is a typical articulated object motion. The most critical characteristics of the articulated objects is that the motion of every body part can be transformed to the global coordinate system by the concatenation of local coordinate systems:

$$X_{global} = T_0 T_1 \ldots\ldots T_{i-1} X_i$$

For example, the global coordinate of a point on the lower arm is equal to the local coordinate of that point multiplied by the transforms of elbow, shoulder and body systems.

A total of 22 degrees of freedom are defined: 6 for the whole body as a free rigid body, 2 for the head, 2 for the upper arms and thighs, 1 for the lower arms and lower legs, 1 for the feet. A typical walking human is shown in Fig.2(c). Including the body part size ratios and 3D model coordinates, there are altogether 58 independent parameters.

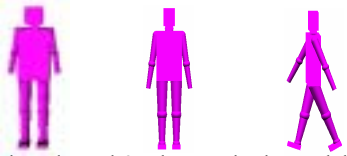

Fig. 2. The adopted 3D human body model: (a) The model in one setting (b) Model with different body part size ratios (c) A walking human

2.3.4.2 Initial parameter setup

With the VO sequence and 3D human model ready, the 2D to 3D inference algorithm is composed of two stages as shown in Fig.3.

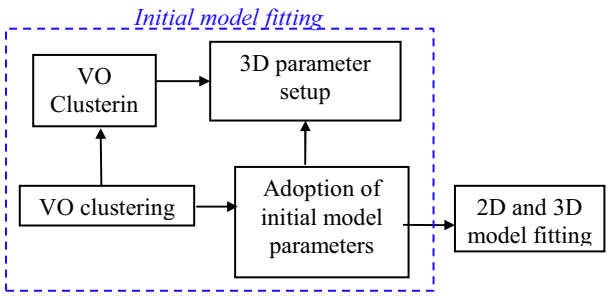

Fig. 3. 2D to 3D Inference for serving shots Algorithm

The first stage is the initial model fitting. The collected VO image frames are divided into two sets: one set is used for training and the other set is used for testing. The training VOs undergo cluster analysis. The features are the seven dimensional Hu's moments and K-means clustering is used. The cluster centers are found and the VO that is closest to the cluster center is identified and considered as cluster center. Through a GUI that is specifically designed to let the user adjust the 3D parameters interactively, initial 3D parameters are manually set up for the cluster centers and kept for further use.

2.3.4.3 3D-2D model fitting

The second stage is 2D and 3D model fitting on the left VO data. Fitting a 3D model to a 2D image shape is a typical multi-dimensional optimization problem. There is a classic solution available for this problem [10]. The algorithm is tailored to be more robust for our application:

1. Given a sequence of tennis player VOs, every VO frame is compared against every cluster center obtained in the initial model fitting stage. The VO which has the shortest distance in feature space to one cluster center is selected as the "seed" frame, which is not necessarily the initial frame in the sequence.
2. An iterative Nelder-Mead's method [14] is used to fit the 3D model to the VO shape in the seed frame. The corresponding cluster center manually crafted 3D parameters are used as the initial guess for the iteration.
3. The estimated 3D parameters obtained in Step 2 for the seed frame are used as the initial guess for the immediate neighboring frames on both sides of the seed frame. The same fitting procedure is used to fit the 3D model to the VO shapes
4. Step 3 is repeated for every frame until all the frames in the sequence are finished.

As the most critical part of the algorithm, the fitting procedure is described as following:

1). A set of camera parameters are assumed. In the tennis games, the shots with interested body shape size are of approximately the same camera setup. So the camera parameters are set to be the same. The camera is assumed to be horizontal and point vertically into the image plane. There is no rotation between the world coordinate system and camera coordinate system. Thus the transformation from world coordinates to camera coordinates is:

$$X_c = X_w + T$$

And the transformation from the camera coordinates to image coordinates is:

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

2). The 3D model is projected to the image plane through the transformation defined in Step 1. The silhouette of the projected 3D model and that of the corresponding VO are exclusively ORed (XOR) [8]. The number of the non-zero pixels is the matching error:

$$E = \sum S_{model} \oplus S_{VO}$$

3). Because we have so many (58 in total) parameters. Directly putting all the parameters into the optimization procedure will easily result in local minimum trapping. So the optimization is performed in a coarse-to-fine style: first the variables related to the whole body including 3D model coordinates, body rotation angles and body size were aligned through the Nelder-Mead's method. Then the remaining variables for the body parts, such as the arm and leg angles, were fine-tuned by the same method. This procedure can be run many rounds until no more change of matching error E being found or maximum iterations being exceeded.

3. EXPERIMENTAL RESULTS

A total of approximately 250 minutes of tennis games were recorded and digitized to 352x240 MPEG1 format from Wimbledon 2004 tennis games. Four separate games were obtained with 8 players including both males and females. The video sequences are segmented into 1344 shots. The removal of audience related shots are very effective: 115 out of 121 audience shots are detected. The coarse level HMM classification results for BS and FF shots are shown in Table 1.

Table 1: Confusion matrix for BS and FF classification

	BS	FF
BS	682	119
FF	30	392

As seen in Table 1, while the classification for FF shot are pretty accurate, the BS shots have a higher mis-classification rate. This is due to the size of the body shapes varies greatly across the BS shots: the BS shots with smaller player body shapes are more

likely to be classified as FF shots with strong motion. Fortunately, the BS shots with small player body shapes rarely contain serving poses, hence the influence on the fine-level analysis is relatively minor.

Due to the limitation of the size of this paper, only the final 2D to 3D fitting results in fine-level analysis is presented here. As shown in Fig.4, (a) shows a VO in serving; (b) shows the corresponding 3D model of the cluster center; and (c) is the final fitted model; (d) shows the initial matching; (e) shows the fitting after the coarse level fitting. It can be seen that the arms of the model is remarkably away from those of the VO. Finally, (f) shows the converged fitting result. Note the change of the body size and the adjustment of the arms. There is only one coarse-to-fine round of model fitting necessary for this VO as discussed in the Step 3 of the fitting procedure–it is true for most of the VOs. The matching error as the iteration progresses is shown in Fig.5. The algorithm is stable: the average parameter estimation difference (over 10 times) is only 2.93%.

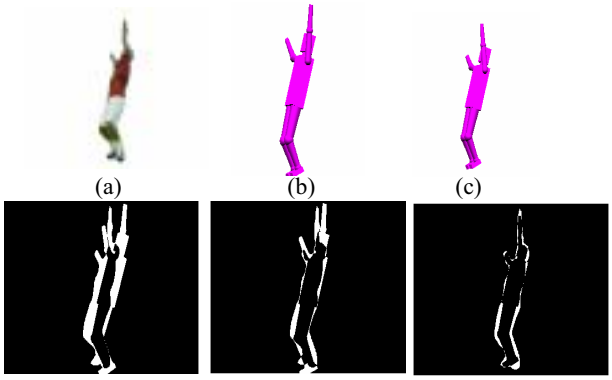

Fig.4. The final fitting results of a serving pose.

4. CONCLUSION AND DISCUSSION

In this paper an innovative 2D to 3D human body modeling fitting algorithm is proposed, based on a comprehensive coarse-to-fine video sequence analysis framework. The proposed scheme is shown to be effective and yield promising results on the demonstrated real world examples. The paper only employs the single-direction knowledge flow from coarse-level to fine-level. It is worthwhile to pursue the feedback from fine-level to coarse-level to improve the accuracy of the appropriate shot detection, thus further increases the robustness of the 2D to 3D parameter inference.

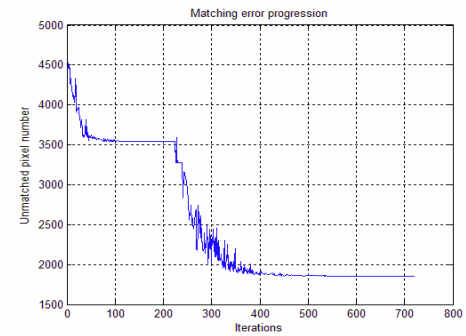

Fig.5. The matching error as iteration progresses

Reference:

[1] F. Cheng; W.J. Christmas, J. Kittler, Recognising human running behaviour in sports video sequences. *ICPR'02,* Québec City, Canada, Aug. 2002.

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

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

[4] I. Karliga and J.-N. Hwang, A framework for fully automatic moving video-object segmentation based on graph partitioning and object tracking*, IEEE MMSP'04*, Siena, Italy, 2004.

[5] C. Kim and J.-N. Hwang, Video Object Extraction for Object-Oriented Applications. *Journal of VLSI Signal Processing*, 29(1/2):7-22, Aug., 2001.

[6] C. Kim and J.-N. Hwang, Fast and Automatic Video Object Segmentation and Tracking for Content-Based Applications. *IEEE Tran. CSVT*, 12(2):122-129, Feb. 2002.

[7] D. Koubaroulis, J. Matas, J. Kittler, Colour-based object recognition for video annotation. *ICPR'02,* Québec City, Canada, Aug. 2002..

[8] H. Lensch, W. Heidrich, H.-P. Seidel, A silhouette-based algorithm for texture registration and stitching. *Graphical Models* 63, 245-262 (2001).

[9] T. Lin, H.-J. Zhang, Automatic video scene extraction by shot grouping. *ICPR'00,* Barcelona, Spain, Sept. 2000.

[10] D. Lowe, Fitting Parameterized Three-Dimensional Models to Images, *IEEE Trans. PAMI*, 13(5):441-450, May 1991.

[11] H. Lu, Y.-P. Tan, Sports video analysis and structuring. *Proc. 2001 IEEE Fourth Workshop Multimedia Signal Processing,* Cannes, France, Oct. 2001.

[12] Y. Luo, J.N. Hwang, Vido sequence modeling by dynamic Bayesian networks: a systematic approach from coarse-to-fine grains. *ICIP'03*, Barcelona, Spain, Sept. 2003.

[13] H. Miyamori, S.-I. Iisaku, Video annotation for content-based retrieval using human behavior analysis and domain knowledge. *ICAFGR'00,* Grenoble, France, Mar.2000.

[14] J. A. Nelder, R. Mead, A simplex method for function minimization. *Comput. J.* 7, 308-313, 1965.

[15] M. Petkovic, V. Mihajlovic, W. Jonker, Techniques for automatic video content derivation. *ICIP'03,* Barcelona, Spain, Sept. 2003.

[16] A. Sappa, N. Aifanti, S. Malassiotis, M. Strintzis, Monocular 3D human body reconstruction towards depth augmentation of television sequences, *ICIP'03,* Barcelona, Spain, Sept. 2003.

[17] S. Wachter, H.-H. Nagel, Tracking Persons in Monocular Image Sequences, *CVIU* 74(3):174-192, 1999.