Real-Time Video Content Analysis Tool for Consumer Media Storage System

Jungong Han, Dirk Farin, Peter H.N. de With, *Senior Member, IEEE,* and Weilun Lao

Abstract — *With the growing storage capacities of hard-disks and optical discs, large consumer video databases are gradually developing. The large effective storage capacity using compressed video leads to the application of fast storage and retrieval functions, to enable quick user-friendly searching for and access to specific parts of the video data. We explore the feasibility of a near real-time semantic sports video analyzer for an experimental consumer media server. This tool is able to automatically extract and analyze key events in a video sequence. The analyzer employs several visual cues and a model for real-world coordinates, so that key parameters (e.g. speed and position) of a player can be determined with sufficient accuracy. This special data can be stored as metadata, thereby facilitating intelligent searching of events. The tool consists of four processing steps: (1) playing frame detection, (2) court extraction, as well as a 3-D camera model, (3) player segmentation and tracking, and (4) event-based high-level analysis exploiting visual cues extracted in the real-world. Our system has been evaluated in a new distributed AV content analysis system for home entertainment. We show attractive experimental results indicating the system efficiency and classification skills, thereby offering new analysis and search/retrieval tools to the consumer.*[1]

Index Terms — Media server, analysis tool, sports video, content analysis, semantic, feature extraction, real-time.

I. INTRODUCTION

With the advent of hard-disk video recording and emerging large-capacity optical disc technology, such as HD-DVD and Blu-ray disc, large video databases for consumer applications are gradually developing. The large storage capacity of compressed video on disks increases the need for fast storage and retrieval functions, realizing quick user-friendly searching for and access to specific parts of the video data. The identification of those parts in the video can be improved or enhanced by metadata, describing the specific properties of key objects in the video scene. Such metadata should then be generated by a tool, which is active at recording time. This paper explores such a tool for finding rich content and presents an experimental architecture for home entertainment, in which it can be integrated.

[1] This work was supported by the ITEA (Information Technology for European Advancement) project Candela, featuring industry and academia.

Jungong Han, Dirk Farin and Weilun Lao are with Eindhoven University of Technology, Den Dolech 2, 5600MB, Eindhoven, The Netherlands (e-mail: {jg.han, d.s.farin, w.lao}@tue.nl).

Peter H.N. de With is with Eindhoven University of Technology, Signal Processing Systems Group, He is also with LogicaCMG Eindhoven, 5605JB, Eindhoven, The Netherlands (e-mail: P.H.N.de.With@tue.nl).

In consumer television, sports videos constitute a major percentage of the total of video content that are provided by public and commercial television channels. Because of the growing offer of TV channels providing full coverage of large sports events, we have focused on sports video analysis and finding meaningful parameters and event data.

A. State-of-the-Art of sports video analysis

Previous approaches to the semantic analysis of sports video can broadly be classified into two genres. Earlier publications [1]-[3] attempt to exploit the general features to extract highlights from the sports video. In [1], the replay action is used to indicate which segment is emphasized by the director so that it should be the highlight of the game. In this way, these systems [1], [2] merely detect the replay segments in sports video and directly label them as the highlights. In [3], authors propose to use heuristic audio cues like vocal reactions of audiences and volume of the commentator, to indicate the content of the sports game. For example, in a soccer game, the cheers of the audience or the sudden increase of the commentator's volume, leads to the assumption that an important event like a goal shooting occurs. Obviously, some of the above algorithms can be applied to analyze various sports games, because those features are more general. However, this kind of analysis scheme is unable to provide sufficient understanding of a sports game, since a viewer cannot deduce the whole story from looking to a special event only. In other words, there is still a big gap between highlight detection and important event *recognition*. The second genre of semantic analysis systems aims at using the domain knowledge of particular games to infer the events [4]-[7]. In [4], the positions of players and the court (playing-field) are utilized to detect a free-kick. In the following, we will focus on tennis videos in order to restrict ourselves in terms of the analysis space. However, some of the concepts can be applied to other sports as well, such as playing-field detection and player tracking. Sudhir *et al.* [5] propose the first tennis video analysis system approaching a video retrieval application. It detects the court lines and tracks the moving players, after which it extracts the events, such as base-line rally, based on the relative position between the player and the court lines. Unfortunately, the proposed scene-level analysis model is rather limited, because only position information is employed to extract events. The system described in [6] is an improved version of [5], but some problems like court-line detection and scene-level model are still not completely resolved. Kijak *et*

Contributed Paper
Manuscript received July 15, 2006

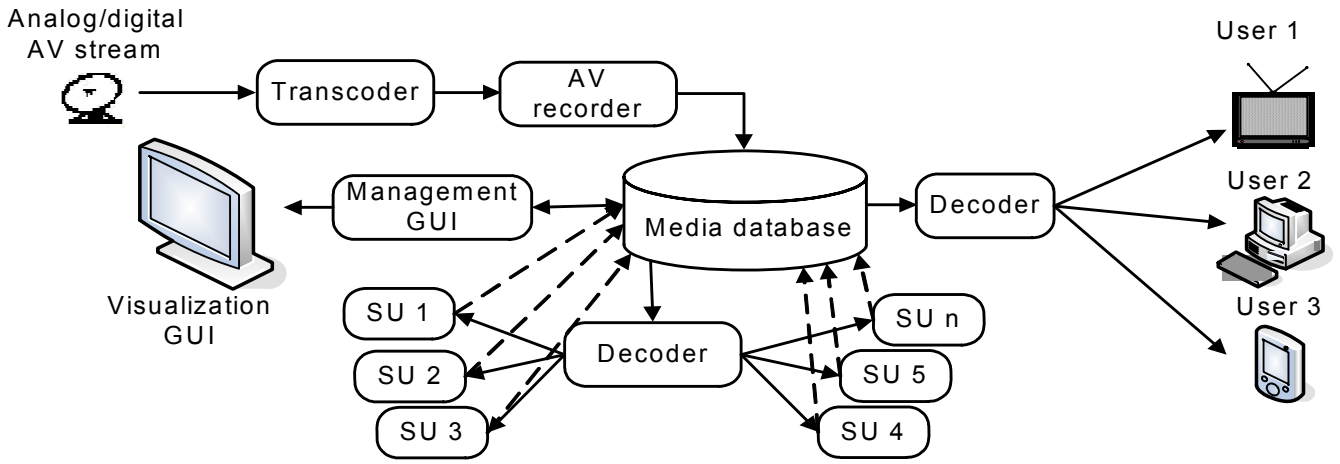


Fig. 1. Real-time distributed networked system requiring multimedia content analysis using service units for face detection, speech recognition, etc.

al. [7] first define four types of views in tennis video, involving global, medium, close-up and audience, and then detect events like first-service failure in terms of the interleaving relations of these four views. This shot-based model does not take object behavior into account, so that it is impossible to provide sufficient classification capabilities.

B. Requirements of home-use sports analysis systems

Sports analysis in consumer applications offers potential interests for the users of the system. However, although many sports video processing algorithms aiming at different sports types have been proposed, sports analysis for a home-entertainment system is not easily achieved, because the requirements are different for each user. The specific challenges for consumer applications are as follows.

1. A broad range of different analysis results that can facilitate various users.
2. A system that provides analysis at different semantic levels. This involves that the gaps between the pixel, object-level and scene-level are bridged in a joint analysis tool.
3. High processing efficiency achieving (near) real-time operation with low-cost consumer hardware.

In this paper, we present a fully *automatic* and *real-time* system towards multi-level analysis of tennis video sequences, belonging to the second genre introduced above. The main contributions of our system are in two aspects. Firstly, we make use of a 3-D camera model to bridge the pixel, object-level and scene-level of tennis sports analysis, which enables the provision of various semantic results for different users. Secondly, a *weighted* linear model combining the visual cues in the real-world domain is proposed to identify events, since the importance of each visual cue to different events is not exactly equal. Adaptive adjustment of weight factors for each visual cue to different events ensures that our algorithm achieves a high accuracy. With this simple but efficient linear model, our algorithm can be used for realizing a *real-time* system, which is difficult to obtain with the alternative learning-based event classification methods. Our system is capable of classifying service, base-line rally and net-approach events, which are consistent with the viewer's understanding about a tennis game. Furthermore, our semantic analyzer was embedded in a new experimental *real-time* and *distributed* AV content-analysis system, as illustrated in Fig. 1. This setup was developed in cooperation with the industry [8]. More specifically, this AV content-analysis framework includes advanced analysis components (Service Unit SU) such as audio classification, automatic speech recognition, audiovisual scene segmentation, sports video analysis and face recognition, connecting via a network. The sports video files (MPEG-2 format) are stored in the Media Database (MDB), and streamed to our analysis unit. The resulting XML data of our unit is streamed to the real-time visualization GUI application and to the MDB unit for storage, and can also be transmitted to other SUs for reuse. This distributed content analysis within the above networked system was successfully demonstrated at an international multimedia conference [8].

In the sequel, we first present a system overview in Section 2 and then describe in detail the semantic analysis module in Section 3. The experimental results on tennis video sequences are provided in Section 4. Finally, Section 5 draws conclusions.

II. OVERVIEW OF PROPOSED TENNIS VIDEO ANALYSIS SYSTEM

Fig. 2 shows the architecture of the proposed analysis system, which consists of four modules, each being briefly explained below.

A. Playing-Frame Detection Based on White-Pixel Ratio

A tennis sequence not only includes scenes in which the actual play takes place, but also breaks or advertisements. Since only the playing frames are important for the subsequent processing, we efficiently extract the frames showing court scenes for further analysis. In our system, the playing-frame detection only identifies the white pixels of court lines and distinguishes the difference between the numbers of white pixels inside two consecutive frames. We

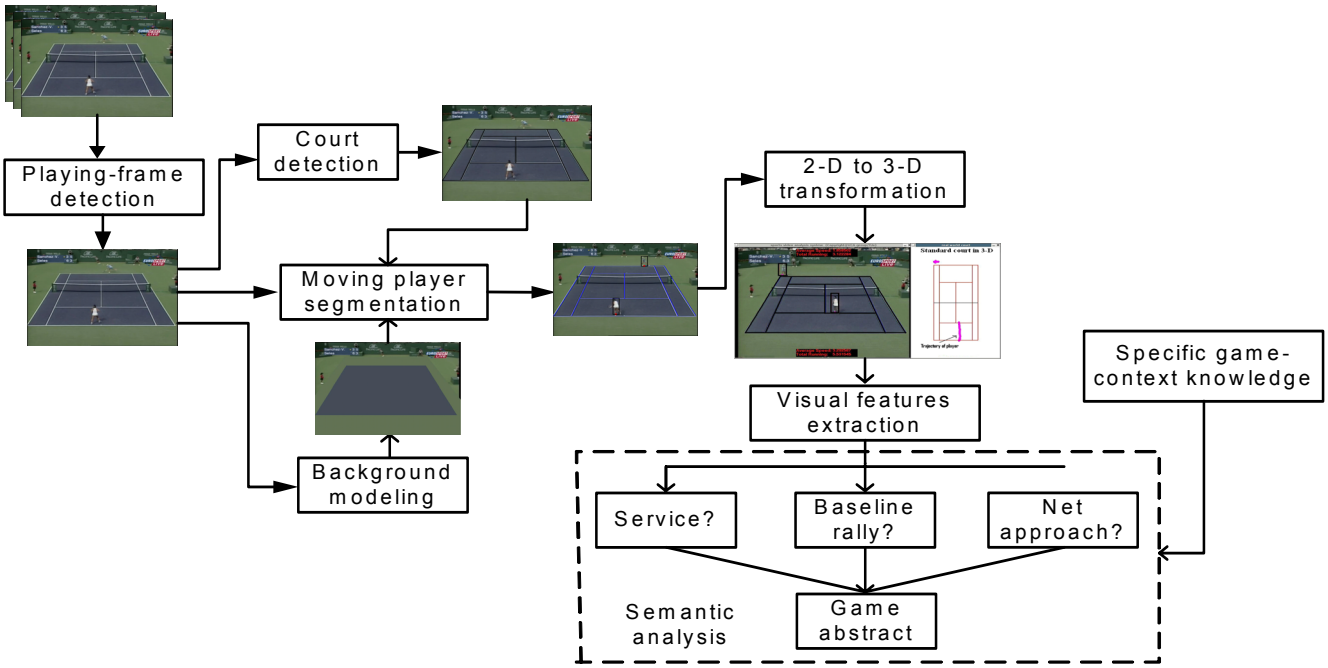


Fig. 2. Architecture of the complete sports video analysis system with small sample pictures showing the intermediate results after each stage.

use this metric [9], because we found that the color of the court line is always white, irrespective of the court type, and the number of white pixels composing the court lines is relatively constant over a large interval of frames (several hundreds). Compared to conventional techniques [5], [6] based on the *mean* value of the dominant color, this technique is more efficient and removes a complex procedure for training data.

B. Court Detection and Camera Calibration

Court information, including size, shape and location, is an important aid to analyze the tennis game. To deduce the semantic meaning from the position and movements of the players, their position has to be known in real-world coordinates. However, pixel-level image-processing algorithms will only give the player positions in image coordinates, which are physically meaningless. To transform these image coordinates to physical positions, a camera-calibration algorithm has to be applied [10].

The task of a camera-calibration system is to provide the geometric transformation that enables to map points in the image to real-world coordinates on the sports court. Since both the court and the displayed image are planar, this mapping is a homography, which can be written as a 3×3 transformation matrix H. This matrix transforms a point P, denoted as a vector $P=(x,y,w)^T$, from real-world coordinates to image coordinates $P'=(x',y',w')^T$. The transformation is performed by computing $P'=HP$. The transformation matrix H can be calculated from four points whose positions are both known in the court model and in the image. However, because the direct detection of point features is not robust against occlusion, we instead use the intersection points of the court lines. The complete camera-calibration system comprises the following algorithmic steps.

- **Court-line pixel detection.** This step identifies the pixels that belong to court lines. Since court lines are usually white, this step is essentially a white-pixel detector. The mandatory feature of this step is that white pixels that do not belong to court lines (e.g. the player's white clothing, etc.) should not be selected.
- **Line-Parameter Estimation.** Starting with the detected white pixels, line parameters are extracted. We apply a RANSAC-based line detector, which hypothesizes a line using two randomly selected points. If the hypothesis is verified, all points along the line are removed and the algorithm is repeated to extract the remaining dominant lines in the image.
- **Model Fitting.** After a set of lines has been extracted from the image, we need to know the line correspondences between the image and the court model. This assignment is obtained by a combinatorial optimization in which different configurations are evaluated.
- **Court Tracking.** When the initial position of the court is known, the computation in successive frames can be carried out more efficiently, since the position of the court in the successive frame will be close to the previous position.

At the algorithm start and after shot boundaries, Steps 1-3 are carried out to find the initial location of the court in the first image. For the subsequent frames, only Steps 1 and 4 are applied, as they are both computationally inexpensive, so that a high tracking speed is obtained.

C. Moving-Player Segmentation

To analyze a tennis video at a higher semantic level, it is necessary to know where the players are positioned. Earlier systems propose several moving-player segmentation algorithms. A class of methods is based on motion detection [5], [11], in which subtraction of consecutive frames is followed by applying a threshold to extract the regions of motion. Obviously, with such a simple detection algorithm, it is impossible to analyze cases where the background is also moving or the camera is moving at the same time. Another category proposes the use of change-detection algorithms. In change-detection systems, the background is first constructed, and subsequently, the foreground objects are found by comparing the background frame with the current video frame. The literature addressing tennis analysis [12], [13] concentrates on selecting a video frame of the tennis court without any players as a background image and then segmenting the players in the video sequence by looking for variations within the background. Unfortunately, in most tennis videos, such frames rarely occur. In conclusion, earlier systems adopt existing techniques of moving-object detection without any exploitation of specific properties of the tennis video game, which leads to a poor detection performance. In addition, the purpose of detecting players is to obtain the player's positions. That is, only the feet positions of the players are really important for further analysis, but there is no technique addressing this feature specifically.

The contribution of our technique is also based on change detection, but we focus on building a high-quality background based on the game properties of tennis, since the performance of the change-detection technique largely depends on the quality of the background. We have found that in most tennis video sequences, a regular frame containing the playing field mainly includes three parts: (1) the court (playing-field inside the court lines), (2) the area surrounding the court and (3) the area of the audience. Normally, the moving area of the players is limited to the field inside the court and partially the surrounding area. Moreover, the color of the court is uniform, as is also the case for the surrounding area. The above observations have been exploited to separately construct background models for the field inside the court and the surrounding area, instead of creating a complete background for the whole image. Using this concept, two advantages occur as compared to the conventional algorithms. First, a background picture with better quality is obtained, which cannot be influenced by camera motion. Second, because of the improved background picture quality, only color and spatial information are considered for further feature extraction, which makes our proposal simpler than advanced motion-estimation methods. More details about the algorithm can be found in [9].

D. Scene-level event classification

The semantic inference module is based on the definition of a set of events. For the consumer, an event would be an important mark in the play, a fault, etc. For the analysis, events are defined by a linear combination of a number of real-world visual cues, such as instant speed of each player, speed change of each player, distance of the moving players to a set of reference locations (base-line, service-line) and temporal relations between each event. Moreover, we have also found that the importance of each visual cue to different events is not exactly equal. For example, the position of the player is evidently more important to identify a net-approach event than other visual cues. Thus, we propose to assign a weight factor to each visual cue, whose value is depending on its importance to a specific event. Using such a refined weighted linear combination has the potential to yield a higher accuracy. This definition of an event has the advantage of being flexible, since any event of any time scale can be represented. Afterwards, event recognition is achieved by computing a likelihood degree, which also provides a reliability indication of the event recognition.

III. SEMANTIC INFERENCE

The *semantic inference* module derives several real-world visual cues from the image domain and afterwards makes weighted models for event recognition. To achieve the first part of this task, the system should correctly bridge the gap between the numerical image features of moving players and symbolic description of the scene. To do so, we first analyze the game rules and select a list of several visual cues that really facilitate semantic analysis of the tennis game. Second, we try to describe each key event making use of the selected visual cues from a real-world viewpoint, and further analyze which cue is more important to a specific event. The previous two steps are performed off-line, and yield mapping and computing models for events. These models are used in the algorithms for on-line computations of both steps. Third, we compute a likelihood degree of each event for each input frame and decide on the mapping of input frames to events. At the end of this step, a simple but efficient temporal filter is used to extract the start time and the end time of each event. Fourth and finally, we summarize the game based on temporal correlations among events.

A. Real-World Visual Cues in Tennis Video

As mentioned earlier, some existing tennis-video analysis systems [12], [14] employ two common visual features: position and speed of the player. In this paper, we not only extend these two cues to the real-world domain, but also propose two novel cues for event identification in tennis video, which makes it possible to detect more events. Let us now list and motivate the four real-world visual cues that are used by our algorithm.

- **Instant Speed of the Player:** The speed of each player is definitely important, because it reveals the current status of a player (running or still) and it also indicates the intensity of the match.
- **Speed Change of the Player:** Acceleration and deceleration of a player occurs during changes in action behavior.
- **Relative Position of the Player to the Court Field:** This position is instrumental for the recognition of

those events that are characterized by a typical arrangement of players on the playing field.

- **Temporal Relations among Each Event**: In some sports games like tennis and baseball, there are strong temporal correlations among key events. For example, in a tennis video, service is always at the beginning of a playing event, while the base-line rally may interlace with net-approaches.

B. Visual-based Model for Each Event

With the above real-world cues, we can model key events, of which three are given below.

- **Service:** This event normally starts at the beginning of a playing event, where two players are standing on the opposite half court, and where one is at the left part of the court, and the other is at the right part. In addition, the receiving player has limited movement during the service.
- **Base-line Rally:** This occurs usually after the service, where two players are moving along their base-lines with relative smooth speeds, that is, there is no drastic speed change.
- **Net-approach:** This is one of the highlight parts of a game, in which standard visual cues are a large speed change combined with close positioning of players to the net lines.

All event models utilize the four real-world visual cues described earlier. We have found that a linear combination of these visual features can be applied to identify the events. Furthermore, we utilized that the importance of each cue to different events is not equal. For example, the temporal position is clearly more important than other cues in order to identify a service event, as 90% of the frames belong to the service event within the first four seconds of a playing event (verified by our sequences). Similarly, the position of the player is more important than other cues to recognize a base-line rally or a net-approach. Therefore, we assign weighting factors to each visual cue and then make linear combinations of the four visual cues.

As an example, we discuss the service-event detection in more detail and illustrate the computation of a likelihood degree for an input frame. The likelihood degree L_i is obtained by

$$L_i = w_1 \times t_i + w_2 \times s_i + w_3 \times a_i + w_4 \times p_i, \quad (1)$$

where i is the frame number, t_i denotes a temporal relation. For service detection, if the current frame is within the first four seconds of a playing event, then $t_i = 1$ otherwise $t_i = 0$. The parameter s_i represents the instant speed of a player. In this case, if the speeds of both players are less than 1 m/s, then parameter $s_i = 1$, otherwise $s_i = 0$. The parameter a_i refers to the speed change of a player. In the service case, if the speed changes of both players are less than 1 m/s, then $a_i = 1$, otherwise $a_i = 0$. Parameter p_i means the relative position between the player and the court field. In the service case, if two players have positions close to the baselines and also on the opposite half court, then $p_i = 1$, otherwise $p_i = 0$. Weighting factors w_1, w_2, w_3 and w_4 correspond to each feature. In the service case, $w_1 = 2$, but w_2, w_3, and w_4 all equal unity, as temporal relations are more important than other features. In our algorithm, when $L_i = 3$, we mark this frame as a service frame. Similarly, this formula can also be used to extract the base-line rally and net-approach by merely changing the variable values and weighting factors related to a different event model, which makes this likelihood concept generally applicable.

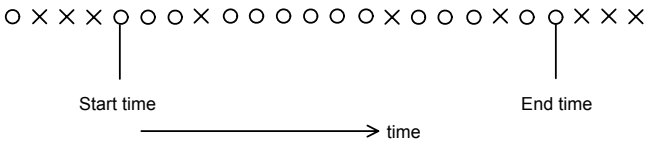


Fig. 3. An example showing how to extract the start time and the end time of the service event.

C. Event Extraction

Up to now, each frame is event-classified as a service, base-line rally or net-approach. The next step is to extract the start time and the end time of each event. Fig. 3 portrays a practical example, where we show a set of frames in the temporal direction. A circle represents a detected "service" frame, and a cross stands for a "non-service" frame. It can be noted from Fig. 3 that the first frame is not a suitable start frame of the service event, although it is labeled as a "service" frame. This is because there are three "non-service" frames behind it, so that the probability that it is erroneously classified as "service" frame is large. Therefore, the first step of the start-time extraction is to measure the local correlation of the i^{th} frame using the following definition:

$$c(i) = s(i-2) + s(i-1) + s(i) + s(i+1) + s(i+2), \quad (2)$$

where i is the frame number, and $s(i)$ denotes the binary state of this frame. If frame i is a service frame, $s(i) = 1$, otherwise s(i) equals 0. We compute the local temporal dependence $c(i)$ for each "service" frame (circles), then select the lowest frame number i for which $c(i) > 2$ as the start frame of the service event.

In order to detect the last frame of the service, we first calculate the occupancy rate by

$$O_i = n_{j,i} / N_{j,i}. \quad (3)$$

Here, assuming the current i^{th} frame is classified as "service", $n_{j,i}$ counts the number of "service" frames between the first "service" frame with index j and the current frame i. Furthermore, parameter $N_{j,i}$ denotes the total amount of image frames between the first "service" frame and the current frame (include some frames that are classified as non-service), hence $N_{ij} = i - j + 1$. We compute O_i for each "service" frame. The frame with the largest index i for which $O_i > 0.7$ is selected as the end frame of the service. The threshold value 0.7 was derived after conducting several experiments.

D. Game Summary

Our scene-level analysis not only identifies some important events, but also intends to summarize the game, making use of temporal correlations between events. For instance, if there is a service event without a base-line rally or a net-approach that directly changes to a "non-event", it is reasonable to deduce that such a case might be an ace or a double-fault. Furthermore, it is feasible to calculate how many net-approaches each player carried out during a match. Based on the statistical results, the player with more net-approaches is classified as more aggressive.

IV. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed algorithm, we tested our system using 7 broadcasted tennis video clips (totally more than 40 minutes) recorded from three different tennis matches (US open, Australian open, and French open).

A. Standalone experiments with our sports analysis system

(1) Results for court detection and player segmentation

We evaluated our playing-frame detection algorithm, court-detection algorithm and player segmentation algorithm. The system achieves a 98% detection rate on finding court-view frames, 98% rate on court detection and camera calibration, and 96% detection of players, where the criterion is that at least 70% of the body of the player is included in the detection window. Here, the ground truth data are manually labeled. Figures 4 and 5 portray a set of practical visual detection results. It can be concluded from these results that our proposed algorithm not only accurately segments the player and the court, but also detects the position of the player in the image domain, irrespective of the court type.

TABLE I
CLASSIFICATION RESULTS

		Detect	Correct	Miss	False
Our model	Service	18	16	0	2
	Baseline rally	16	14	0	2
	Net-approach	6	6	0	0
Linear model	Service	15	14	2	1
	Baseline rally	13	12	2	1
	Net-approach	5	5	1	0

(2) Results for Scene-Level Analysis Algorithm

We manually labeled all events to obtain the ground truth. Table 1 shows the results of our simulations using the proposed weighted linear combination model and the conventional linear model. The results clearly show that our model is better than the conventional linear solution. Also, it can be concluded that the scene-level event extraction rate of the system is about 90%.

B. Application embedding in consumer AV analysis system

Our sports analyzer was embedded in an industrial experimental home-entertainment system, namely a *real-time* and *distributed* AV content analysis system, which was successfully demonstrated at the *IEEE international conference on Multimedia & Expo (ICME2005)*. Fig. 6 shows the user interface of our sports-analysis component. From this user interface, the viewer can find analysis results at three different levels, which provides sufficient analysis results for various users with different preferences. At the pixel level, several key objects are segmented and indicated. Meanwhile, the system indicates whether the current frame is a court-view frame or not. At the object level, the moving objects are tracked in the 3-D domain (at the right side). Several useful data parameters are provided, such as the instant speed of each player, the average speed of each player (in meters/second) and the total running distance. At the scene level, the system automatically classifies three important events including service, baseline rally and net-approach.

Fig. 7 portrays an example of a management *Graphical User Interface* (GUI) in the consumer system [8], which allows a developer to visualize discovered Service Units (SUs), to control SU-related parameters, to create connections between SUs and to monitor existing SU-interconnections. In this Figure, each SU is an independent process that communicates with other SUs using TCP/IP for data streaming. Furthermore, all the SUs are connected to the SU Media Database (MDB), a sort of central memory of the framework, which serves as a persistent storage of both content data and acquired and generated metadata. Moreover, SU MDB maintains links between content and content-related metadata, generated by other SUs. Consequentially, the presence of this SU enables delayed, sequential or off-line metadata processing. We have also tested our sports analyzer in this system, achieving a near real-time performance (2-3 frames/s for 720×576 resolution, and 5-7 frames/s for 320×240 resolution, with a P4-3GHz PC). More important is that our analysis results (XML format) are able to generate the metadata facilitating AV-management-related applications such as search and retrieval. The search and retrieval commands are given by the consumer. It goes without saying that user commands need to be translated into visual cues for searching. This translation is still under study.

V. CONCLUSION AND FUTURE WORK

We have proposed a tennis sports analysis system which is intended to be part of a larger consumer media server having analysis features. The analysis system is an aid for the consumer in classifying sports video programs that were recorded in large quantities and stored on the media server. The automatic sports analysis system can generate metadata that can be used for categorizing the video scenes and give support for fast searching and retrieval. The new proposed sports video analysis system features high-level scene analysis

based on real-world visual cues. The main contribution is that a selected list of real-world visual cues is applied to a set of linear-weighted models of individual events. Robustness of event detection is achieved by temporal dependence functions, which produce probabilistic values indicating the reliability of the event occurrence.

In the near future, we will include audio-analysis functions into our sports content-analysis system for an enhanced robustness. Also, the system may be extended to analyze other sport types, like volleyball, badminton and basketball, since several techniques, such as court detection, camera calibration, player segmentation and high-level analysis, are generally applicable.

ACKNOWLEDGMENT

The authors would like to thank our ITEA-Candela partners Philips Research Labs and Bosch Security System Eindhoven for the kind cooperation.

REFERENCES

[1] V. Kobla and D. Doermann, "Detection of slow-motion replays for identify sports videos," *Proc. IEEE 3rd Workshop on Multimedia Signal Processing*, Copenhagen, Denmark, pp. 135-140, 1999.

[2] H. Pan, P. Beek and M. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," *Proc. ICASSP2001*, pp. 1649-1652, Salt Lake City, UT, May 2001.

[3] K. Wan, X. Yan, X. Yu, C. Xu, "Real-time goal-mouth detection in MPEG soccer video," *Proc. ACM Multimedia*, pp.311-314, 2003.

[4] A. Ekin, M. Tekalp and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Circ. Syst. Video Technol.,* Vol. 12, No. 7, pp. 796-807, July 2003.

[5] G. Sudhir, C. Lee and K. Jain, "Automatic classification of tennis video for high-level content-based retrieval," *Proc. IEEE International Workshop on Content Based Access of Image and Video Databases,* pp. 81-90, 1998.

[6] C. Calvo, A. Micarelli and E. Sangineto, "Automatic annotation of tennis video sequences," *Proc. DAGM-symposium*, pp.540-547, Springer, 2002.

[7] E. Kijak, L. Oisel and P. Gros, "Temporal structure analysis of broadcast tennis video using hidden Markov models," *Proc. SPIE Storage and Retrieval for Media Databases 2003*, pp.289-299, January 2003.

[8] J. Nesvadba *et al.,* "Real-time and distributed AV content analysis system for consumer electronics networks," *Proc. IEEE International conference Multimedia Expo (ICME)*, pp. 1549-1552, July 2005.

[9] J. Han, D. Farin, P. H.N. de With, "Multi-level analysis of sports video sequences*," Proc. SPIE multimedia Content Analysis Management and Retrieval,* Vol. 6073, pp. 1-12, San Jose, January 2006.

[10] D. Farin, J. Han and P. H.N. de With, "Fast camera-calibration for the analysis of sports sequences," *Proc. IEEE International conference Multimedia Expo (ICME)*, pp. 482-485, July 2005.

[11] G. Pingali, Y. Jean and I. Carlbom, "Real time tracking for enhanced tennis broadcasts," *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR),* pp.260-265, June 1998.

[12] N. Rea, R. Dahyot and A. Kokaram, "Classification and representation of semantic content in broadcast tennis videos," *Proc. IEEE International Conference Image Processing,* pp. 1204-1207, September 2005.

[13] T. Bloom and P. Bradley, "Player tracking and stroke recognition in tennis video," *Proc. WDIC,* pp. 93-97, June 2003.

[14] S. Chang, D. Zhong and R. Kumar, "Real-time content-based adaptive streaming of sports videos," *Proc. IEEE Workshop Content-based Access to Video Library*, pp.139-143, December 2001.

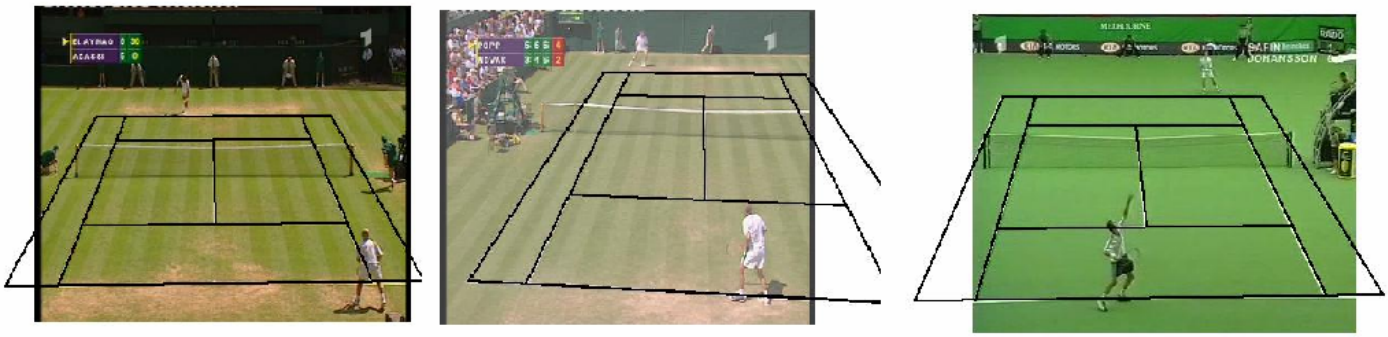


Fig. 4. Court detection results for three different matches. The detected court lines are marked outside the video image to indicate the robustness.

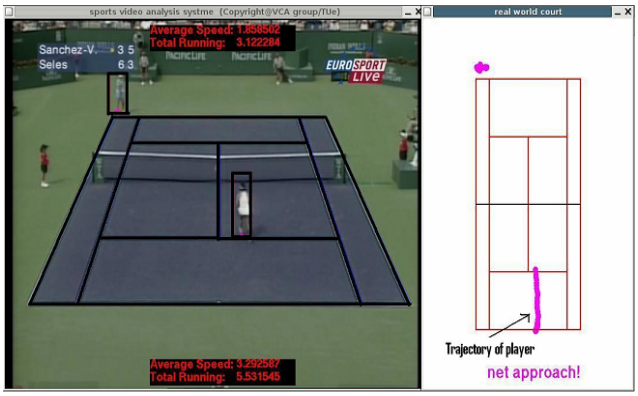


Fig. 6. Results of our analysis system. The left image shows the detected court and players, as well as the average speed of each player (top and bottom). At the right is the real-word court model, where the trajectory of each player is visualized, in this case showing a net-approach.

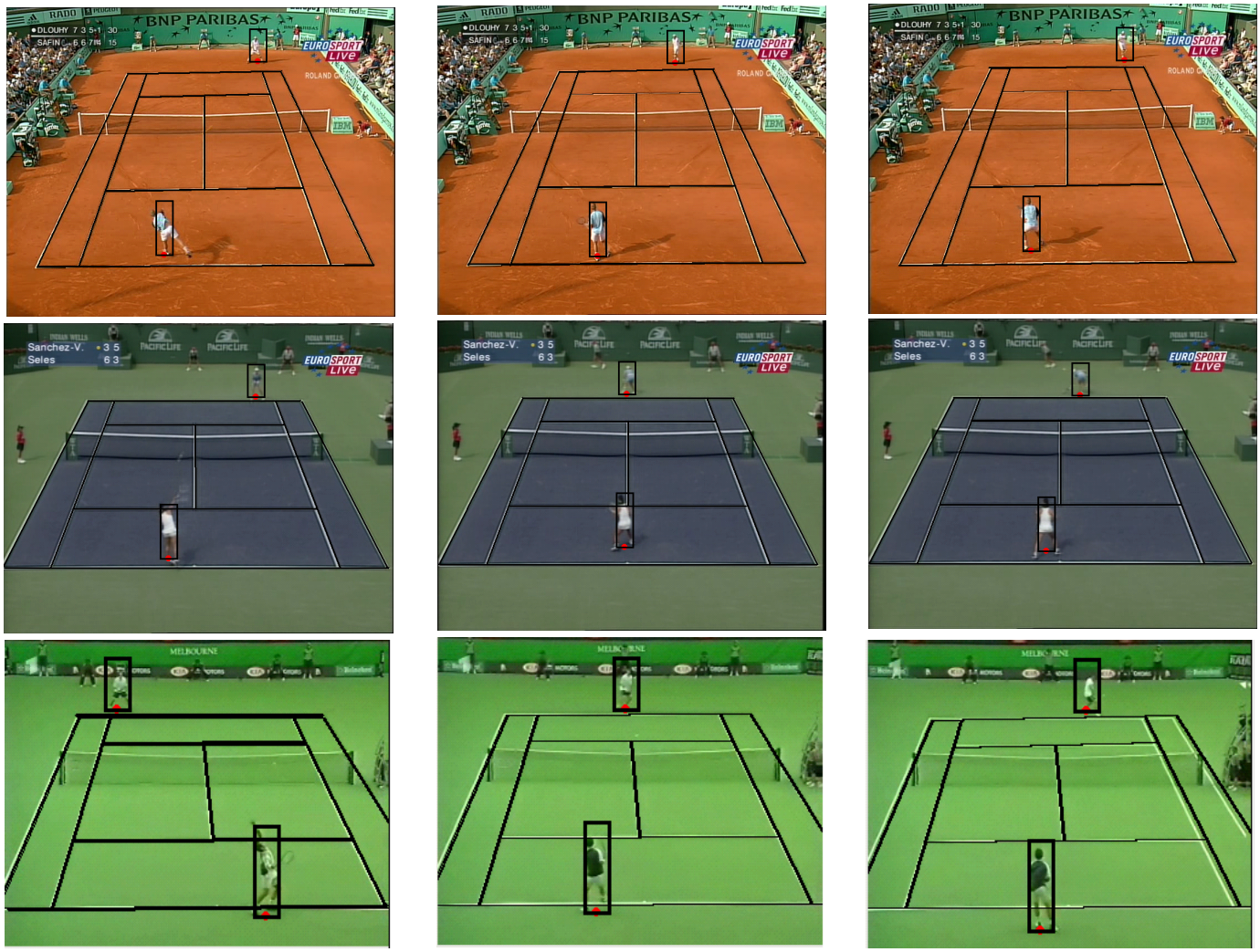


Fig. 5. Player and court-tracking results for 3 consecutive periods of 30 frames (the court is indicated by black lines, the black rectangular represents the detected player).

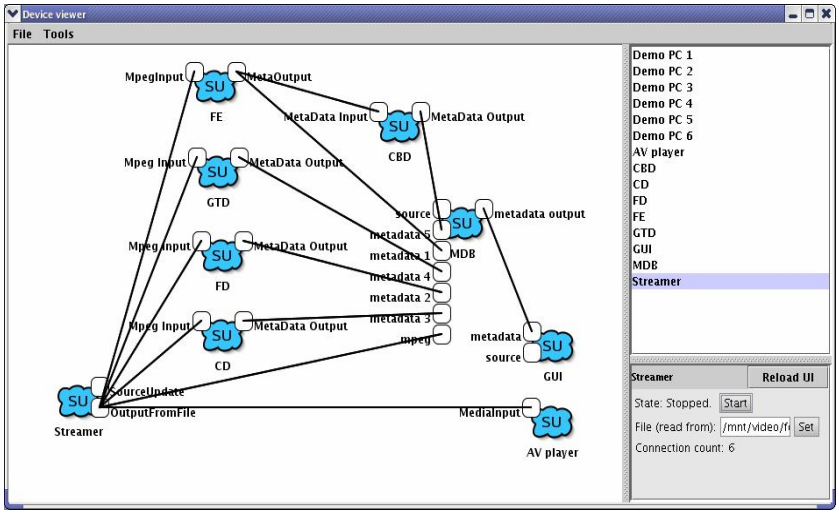
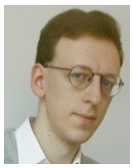


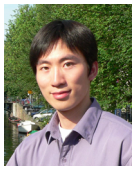
Fig. 7. Management GUI of the ICME 05 demonstration, showing the Service Units with the component names and their relationships.

Jungong Han was born in Xi'an, China, in 1977. He received the B.S. degree in control and measurement engineering from the Xi'an University of Electronic Technology, China, in 1999. In 2004, he received a Ph.D. degree in communication and information engineering from the Xi'an University of Electronic Technology. During 2003, he was visiting student at Internet Media group of Microsoft Research Asia, China, with the topic on scalable video coding. Since 2005, he joined the Video Coding and Architectures (VCA) group at the Technical University of Eindhoven, The Netherlands, where he is currently a Postdoctoral Fellow. His research interests are content-based video analysis, human behavior analysis, image and video compression, scalable video coding and multi-view video coding.



Dirk Farin graduated in computer science and electrical engineering from the University of Stuttgart, Germany. In 1999, he became research assistant at the Department of Circuitry and Simulation at the University of Mannheim. He joined the Department of Computer Science IV at the University of Mannheim in 2001 and the VCA group at the University of Technology Eindhoven in 2004. In 2005, he received his Ph.D. degree from the University of Technology Eindhoven, Netherlands. He received a best student paper award at the SPIE Visual Communications and Image Processing conference in 2004 for his work on multi-sprites, and two best student paper awards at the Symposium on Information Theory in the Benelux in 2001 and 2003. His research interests include video-object segmentation, video compression, content analysis, and 3-D reconstruction. In 2005, he organized a special session about sports-video analysis at the IEEE Int. Conf. on Multimedia and Expo.

Peter H.N. de with graduated in electrical engineering from the University of Technology in Eindhoven. In 1992, he received his Ph.D. degree from the University of Technology Delft, The Netherlands, for his work on video bit-rate reduction for recording applications. He joined Philips Research Labs Eindhoven in 1984, where he became a member of the Magnetic Recording Systems Department. From 1985 to 1993, he was involved in several European projects on SDTV and HDTV recording. In this period, he contributed as a principal coding expert to the DV standardization for digital camcording. In 1994, he became a member of the TV Systems group at Philips Research Eindhoven, where he was leading the design of advanced programmable video architectures. In 1996, he became senior TV systems architect and in 1997, he was appointed as full professor at the University of Mannheim, Germany, at the faculty Computer Engineering. In Mannheim he was heading the chair on Digital Circuitry and Simulation with the emphasis on video systems. Since 2000, he is with LogicaCMG in Eindhoven as a principal consultant and he is professor at the University of Technology Eindhoven, at the faculty of Electrical Engineering. He has written and co-authored over 150 papers on video coding, architectures and their realization. Regularly, he is a teacher of the Philips Technical Training Centre and for other post-academic courses. In 1995 and 2000, he co-authored papers that received the IEEE CES Transactions Paper Award, and in 2004, the VCIP Best Paper Award. In 1996, he obtained a company Invention Award. In 1997, Philips received the ITVA Award for its contributions to the DV standard. Mr. de With is a senior member of the IEEE, program committee member of the IEEE CES, ICIP and VCIP, chairman of the Benelux community for Information and Communication Theory, co-editor of the historical book of this community, scientific board member of LogicaCMG, scientific advisor of the Dutch Imaging school ASCII, IEEE ISCE and board member of various working groups.



Weilun Lao obtained his Bachelor in Dept. Automation from the South China University of Technology (SCUT), China in 2002. From 2003 to 2005, he undertook research work in Institute for Infocomm Research (I2R), Singapore and won his Master in Electrical & Computer Engineering from the National University of Singapore (NUS). He is currently a PhD candidate at Department of Electrical Engineering, Eindhoven University of Technology (TU/e) in The Netherlands. His research interests include multimedia content analysis, 3D human motion analysis and computer vision.