Lawn Tennis Video Summarization based on Audiovisual and Text Feature Analysis

Sudhir S. Kanade
College of Engineering
Osmanabad - 413501 (M.S.), India

Pradeep M. Patil
Sinhgad Technical Institutes Campus Warje,
Pune-48, India

ABSTRACT
In this paper, a new video summarization approach for lawn tennis video is presented. The proposed method uses frame color histogram to classify video into play field color shots (PFCS) and non play field color shots (NPFCS). Play field color shots are the segments of interest and used to recognize the tournament class. A dominant colored frame from every PFCS is extracted as a salient frame. Our approach also employs dominant value of PFCS volume and energy. For the each PFCS dominant audio energy value is computed and corresponding key frame is extracted. On-screen text presented in play field color shots provides important semantic information so the last frame of the shot is extracted as key caption. Three key frames based on audio, visual and text information conveyed in every PFCS are extracted and summary of lawn tennis video is created. These key frames of every PFCS serve as accurate localization of the events on the lawn tennis field. Experiments are performed on lawn tennis videos to confirm the efficiency of the proposed method.

General Terms
Shot boundary detection, video summarization

Keywords
Shot Classification, Color histogram, key frame, play field color shots, non play field color shots, Video Summarization.

1. INTRODUCTION
Sports videos form very significant multimedia content which is viewed globally by large crowd on the internet and mobile devices. Since sports videos are lengthy, most viewers are interested to watch particular segments of sports video of interest to entire video within short time. Watching a live game is exciting, but when it comes to browsing the action that had already taken place, it is a time consuming effort. There is a big challenging problem to efficiently manage sports video contents of important events. The phenomenon inspires us to compress long sequence into a more compact representation through a summarization process.

In this paper, a novel video summarization approach for lawn tennis video is presented. The proposed summarization scheme is different from the presented methods so far used for video summarization. The proposed approach provides a better summarization result for lawn tennis video in terms of the feedback by user study.

Researchers have been using audio activities for video summarization of ball games, news video and in violence detection in movies. Qian Huang and Zhu liu [1] separated news and commercials based on nine acoustic features, text independent anchorperson recognized using GMM and generated abstraction of content for multimedia indexing and retrieval in the context of broadcast news.

Rui Cai et al [2] used HMMs to model highlight sound effects (laughter, applause and cheer) and a log-likelihood scores based method to make final decision. Two new spectral features, sub band spectral flux and Harmonicity Prominence introduced to represent key effects and based on this proposal a Bayesian network based approach to discover the high-level semantics of an auditory context. Xi Shao and Changsheng Xu [3] created musical summary using features linear prediction coefficient, the short-time zero crossing rate and Mel-frequency cepstral coefficients (MFCCs) from audio track and shots are detected and clustered from the visual track. Finally created musical video summary by aligning the music summary and clustered video shots. Baoxin Li and Hao Pan [4] analyzed visual and aural signals for events and non-events to form meaningful indexing points and video summarization is achieved by concatenating the event segments.

Dian Tjondronegoro and Yi-Ping Phoebe Chen [5] localized occurrences of highlights using detection of whistle sound, excitement, and text displays. Chen-Hsiu Huang et al [6] extracted audio and video features and synchronized the input audio and video according to their content related features to form a musical video. Rui Cai et al [7] proposed key audio effects with two new spectral features sub band flux and Harmonicity Prominence.

Color does not only add attractiveness to objects but also give more information, which is used as powerful tool in video summarization, browsing, and retrieval. Spatial feature, color analysis so far is used for variety of applications.

Ekin et al [8] proposed dominant color region detection, shot boundary detection and shot classification algorithms that are robust to variations in the dominant color.

N. Benjamas et al [9] used color histogram comparison to detect shot boundaries. Sandra E.F. deAvila et al [10] generated summaries based on color attributes and visual features. Costas Panagiotakis et al [11] presented a key frame selection algorithm based on three iso-content principles, iso-content distance, iso-content error and iso-content Distortion. In the framework of this work, used the color layout Descriptor (CLD) of MPEG-7 standard to guarantee

interoperability. Kamesh Namuduri [12] used features grass color; and pitch color (sand color) to extract views or states from a cricket video. Kenichi Fujimura et al [13] used color histogram of a shot as color information and discovered important intervals having several color change patterns by using the probability model. Zhonghua Sun et al [14] proposed spatial-temporal color distribution based key frame extraction.

Sports videos also contain textual information describing the scores, time and team or player names. Xi Shao et al [15] proposed method which detects and recognizes lyric captions to analyze music video structure and identify the most salient music part to create the summary of music video.

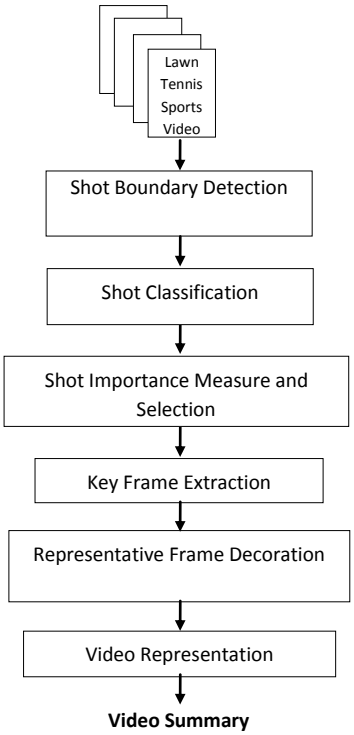


Figure 1. Framework flowchart for Lawn Tennis Video Summarization.

Wonjun Kim and Changick Kim [16] detected overlay text using transition map color-based thresholding method [17] to extract text strings. Cheolkon Jung et al [18] proposed key captions/text extraction method and provided a dual binarization method to segment texts easily with different color polarities from the background.

Most of the above mentioned approaches on video analysis have concentrated on either audio or visual information. Very few has used audiovisual and text information to build video summarization. Consequently, in this paper we propose novel frame work for lawn tennis video analysis and summarization using audio, visual and text features. A flowchart of the proposed framework is shown in Figure 1.

The remainder of the paper is organized as follows: Sections 2 describe the shot importance measure. Sections 3 describe the Audio, Visual and Text feature based salient key frame extraction. Section 4 presents the video summarization. In Section 5, we evaluated the performance of the proposed system. Finally, conclusion is addressed in Section 6.

2. SHOT IMPORTANCE MEASURE

Let us assume a video of N shots $\{S_1, S_2, \ldots\ldots S_N\}$ and each shot contains M frames $\{f_1, f_2, \ldots\ldots\ldots f_M\}$. Shingo Uchihashi and Jonathan Foote [19] proposed a shot is important if it is both long and rare. I-Cheng Chang and Kun-You Chen [20] determined a Shot importance by computing motion energy and color variation. Lawn tennis tournament has one distinct dominant play field color. The dominant color of French open is brown, Wimbledon is green and US open is blue. In the proposed method, Lawn tennis tournament video shots are classified into PFCS and NPFCS. PFCS are the segments of interest and NPFCS are insignificant shots. In Lawn Tennis short PFCS also serve as localization of events like a faulty service and dead a ball. So every short as well as long PFCS are very important. These PFCS not only localize important events but also classify tournament class. So only PFCS are measured by the mean values of each color component from average color histogram of the frames in PFCS. All PFCS are important to classify tournament and serve as accurate localization of events like a faulty service, dead a ball and fall a ball. Furthermore these shots are analyzed to extract three key frames based on audio, visual and text features.

3. AUDIOVISUAL AND TEXT FEATURE ANALYSIS

3.1 Audio Analysis

The aim of using audio information in sports video processing has been either to detect the occurrence of special sound events, such as ball hits, or to locate all interesting events, which are defined either as excited human-speech/crowd segments or high energy segments. Exciting events in sports video are accompanied by high energy segments that may result from crowd noise and/or human speech, the peaks in the audio volume (also loudness or energy) may be sought. Volume does not differentiate the source of the audio segment, but, on the contrary, provides a collective measure, which is desirable in sports video analysis since crowd noise has the same high-level importance as the commentator speech.

In our experiment, all audio streams of lawn tennis are 8-bit, mono-channel, and down sampled to 8-KHz. Each frame is of 256 samples with 50% overlap. The loudness of audio signals is the most prominent feature for human aural perception. Here we use volume and energy to describe the loudness of audio signals. Volume is a key acoustic feature that is correlated to the sample amplitudes within each frame. The volume and energy of each frame is computed as:

A. The sum of absolute samples within each frame

$$Volume1(k) = \sum_{n=0}^{N-1} | x(n) | \qquad (1)$$

Where x(n) is the n-th sample within a frame, and n is the frame size.

B. 10 times the 10-based logarithm of the sum of sample squares:

$$Volume2(k) = 10 * \log \sum_{n=0}^{N-1} x(n)^2 \qquad (2)$$

This method requires more floating-point computations, but it is (more or less) linearly correlated to our perception of loudness of audio signals. The values computed are referred as the "log energy" in the unit of decibels.

C. Audio frame energy is computed for every frame of shot.

$$X(K) = 1/N \sum_{n=0}^{N-1} |x(n)|^2 \qquad (3)$$

From this computed audio frame volume and energy values in the shot, and a dominant frame volume and energy value noted as significant value and the corresponding video frame is extracted as predominant salient key frame in the shot. For each PFCS frame wise audio signal is processed and the predominant frames are extracted as salient key frames for the corresponding shots.

3.2 Visual Analysis

The playing field in lawn tennis tournaments can be described by one distinct dominant color. In lawn tennis video, dominant color of the field decides particular tournament. Like dominant green colored play field is of Wimbledon matches, brown colored play field is of French open tennis and blue colored field is of US open tennis tournaments. For every frame we extract eight channel color histograms in the HSV color space. Color histogram characteristics are used to compute the dominant color in the shot. Based on this dominant color we classify the lawn tennis matches and among these frames selecting one of the dominant colored frame as key frame.

A. Dominant color shot detection using color histogram

Three tournament videos are considered for analysis: 1) French open 2) Wimbledon 3) US open. Every Lawn Tennis Tournament has one diverse dominant play field color (a tone of brown for French open tournament, a tone of green for Wimbledon tournament, and a tone of blue for US open tournament). The statistics of this dominant color in the HSV (hue-saturation-value) space is considered. Hue values are considered for computation of N number of color components. Among N color components red, green and blue color components computed by thresholding hue values. In color histogram analysis red color is treated as French Open Playfield color, green color is treated as Wimbledon Play field color, and blue color is treated as US Open Playfield color. To classify video into play field color shots (PFCS) and non play field color shots (NPFCS) we employed global color histogram of every frame in the shot.

Frame wise color histogram gives the number of times a particular color has occurred in the frame. These histograms are defined as *h1, h2, h3 ... hn* and given by the expression.

$$H(k) = \sum_{i=0}^{N-1} hi \qquad (4)$$

From these values we define mean value of red, green and blue color components of PFCS as

$$H_R(k) = \frac{1}{N} \sum_{i=0}^{N-1} hi \qquad (5)$$

$$H_G(k) = \frac{1}{N} \sum_{i=0}^{N-1} hi \qquad (6)$$

$$H_B(k) = \frac{1}{N} \sum_{i=0}^{N-1} hi \qquad (7)$$

Where, $H_R(k)$ is mean value of red color components, $H_G(k)$ is mean value of green color components, and $H_B(k)$ is mean value of blue color components.

B. Classification of Shots into Tournament Classes

Based on mean value of each color component of shot, Lawn tennis video is classified into relevant tournament class. The segments of interest are play-field colored shots. Motivated by the dominant colored shots we classify PFCS into three tournament classes 1) French Open tournament 2) Wimbledon tournament 3) US Open tournament. The distinctiveness of each tournament class is as given below:

French Open Tournament: The dominant frame color of this class of tournament is a tone of brown. This brown color in histogram computation is considered as red color. Lawn tennis video is classified into this type of tournament class if mean value of red color component is greater than green and blue value i.e. $H_R(k) > H_G(k)$ and $H_R(k) > H_B(k)$.

Wimbledon Tournament: The dominant frame color of this class of tournament is a tone of green, which is uniqueness of this type of tournament. Lawn tennis video is classified into this type of tournament class if mean value of green color component is greater than red and blue value i.e.

$H_G(k) > H_R(k)$ and $H_G(k) > H_B(k)$.

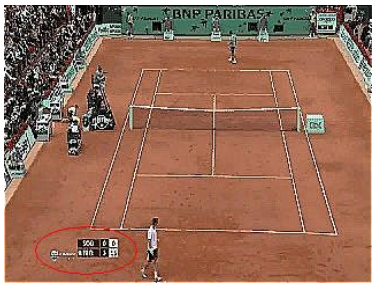
US Open tournament: The dominant frame color of this class of tournament is a tone of blue. Lawn tennis video is classified into this type of tournament class if mean value of blue color component is greater than red and green value i.e. $H_B(k) > H_R(k)$ and $H_B(k) > H_G(k)$.

Classification of a shot into one of the above three tournament classes is based on the mean value of dominant play field color value of shot.

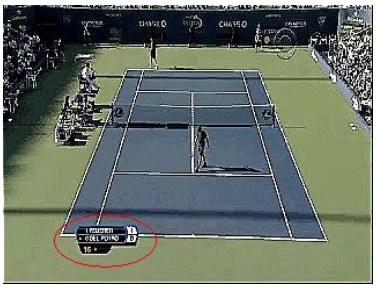
3.3 Text Analysis

Sports videos also contain text describing the scores and team or player names. On-screen text refers to the text overlaid on the video during editing; hence, it comes embedded to the visual information. The definition of on-screen text excludes scene text, which is independent of video editing. An example to on-screen text is shown in Figure 2. Screen text processing has not received much interest in the sports video processing literature.

(a) frame # 338



(b) frame # 148



(c) frame # 708

Figure 2. On screen text examples of Play Field Colored Shots.

In lawn tennis video of Wimbledon tournament the location of the score box with player name and score is generally located at top-left corner for frames of play field shots. Whereas for US open and French open tournament it is located at bottom-left corner of frame. The last frame of the every play field color shot shows the final scores of that shot. So, considered this frame as key caption for this particular shot and extracted as salient key frame for video summarization.

4. VIDEO SUMMARIZATION

A video summary is much shorter than the original video. The video summary is a collection of a set of static representative frames, we preferred three representative frames:

Step 1: The first step of proposed summarization method is to obtain the play field color shots (PFCS) as important shots and non play field color shot (NPFC) as trivial shot.

Step 2: For each PFCS extracted three representative frames which can act as access points or bookmarks. The representative frame is a static representation of the video contents and can be used to express as an outline of the video.

Step 3: First key frame is extracted corresponding to dominant value of PFCS volume and energy.

Step 4: A Second key frame is selected as dominant colored frame from every PFCS This frame shows global view of the field and serves as accurate localization of the events on the field.

Step 5: The last frame of PFCS as key caption frame for the shot.

Step 6: For every PFCS three key frames are selected based on audio, visual and text features.

Step 7: Employed an Unsharp mask filter to decorate extracted representative frames [21].

Step 8: Sequentially layout the key frames of the video shot from top to bottom and left to right.

5. EXPERIMENTAL RESULTS

In this section, the experimental results of the proposed algorithm on various lawn tennis tournament video are presented. The algorithm is implemented using Mat lab on a 2.80 GHz Pentium (R) D computer, running Microsoft Windows XP. The videos of various tournaments are taken from You Tube. We experimented on 05 games (1276 sec) for lawn tennis with 320x240, 400x226, and 720x480 AVI videos. Figure 3 shows the frame sequence of different lawn tennis video PFCS and their color histograms. The color histograms show the dominant play field colors of each frame in the shot. Dominant color is described by mean color; the algorithm estimates the mean values of each color component from average color histograms of the frames in the play field color shot. With this we segmented the video into PFCS and NPFCS. Whereas PFCS are segments of interest and NPFCS are trivial shots. So, average of mean value of frames for red, green and blue color in every PFCS is computed. Dominant color of the shot is evaluated and classified into relevant tournament class. Table 1 shows visual analysis of Lawn

tennis video classification and summarization performance. Here dominant PFCS accuracy and compression ratio is evaluated using

$$\text{Dominant PFCS accuracy } (\%) = 100 \times \frac{\text{Obtained PFCS}}{\text{Desired PFCS}}$$

$$\text{Compression Ratio } = 100 \times \frac{\text{Obtained PFCS timelength}}{\text{Total time Length of game}}$$

From every PFCS three key frames are extracted. First key frame is extracted as dominant colored frame from the PFCS.

Figure 4 and table 2 shows results of audio analysis.

Here volume and energy is computed to describe the loudness of audio signal and extracted second key frame corresponding to dominant audio energy value of the shot.

Figure 2 shows the text analysis of the PFCS. As the last frame of the shot designate the final score of the shot so extracted as third salient frame.

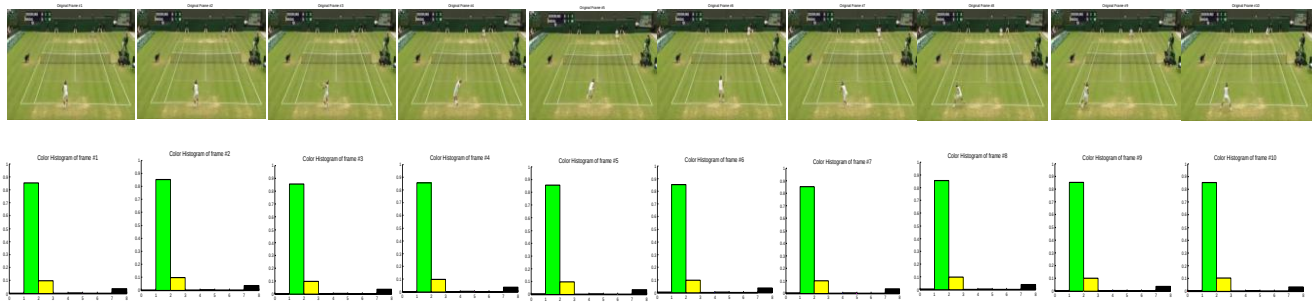
As a result, selected three key frames from every PFCS using audio-visual and text features.

$$\text{Number of key frames extracted per game} = 3 \times \text{btained PFCS}$$

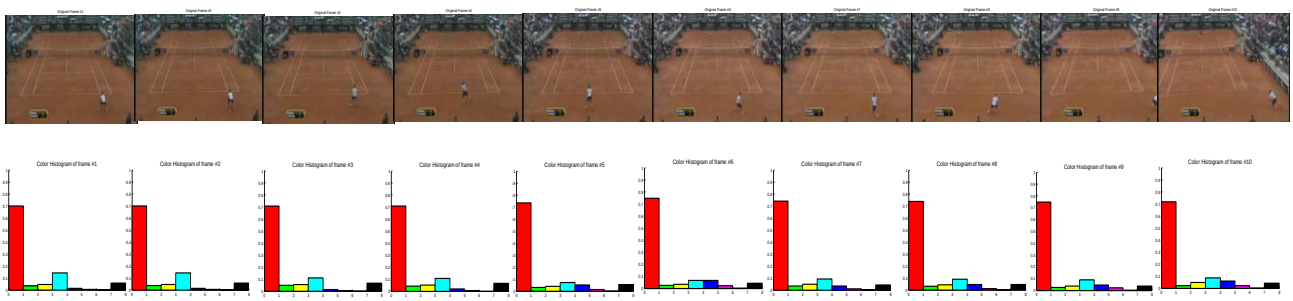
Figure 5 shows the selected and decorated representative frames of French Open, Wimbledon and US Open Lawn Tennis Tournament.

Table 1. Video Classification and Summarization Performance

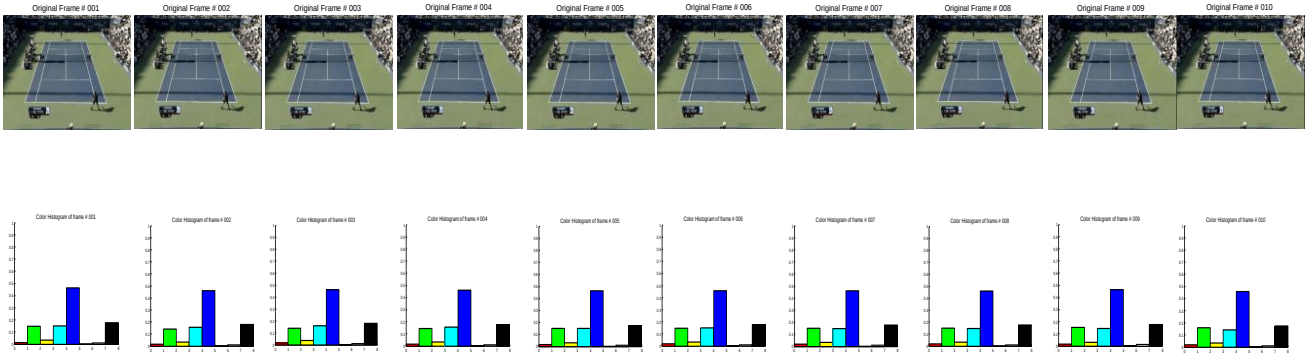
Lawn Tennis Video	Format	Time Length (in sec)	Total Shots	Number of frames	Desired PFCS	Obtained PFCS	NPFCS	Dominant Color of PFCS	Tournament class	Obtained PFCS time length (in sec)	Compression Ratio (%)	Extracted key frames	Dominant PFCS accuracy (%)
Game1	AVI 320x240	381	97	9144	52	42	61	Green	Wimbledon	299	78.47	105	80.76
Game2	AVI 320x240	367	79	8808	59	47	43	Red (Brown)	French open	220	59.94	108	79.66
Game3	AVI 320x240	337	49	8088	40	32	17	Blue	US Open	238	70.62	96	80
Game4	AVI 400x226	50	08	150	06	03	05	Red (Brown)	French open	40	80	24	50
Game5	AVI 720x480	141	27	3384	24	15	12	Green	Wimbledon	115	81.56	45	62.5



(a) Game 1: Example of PFCS frames and their color histograms

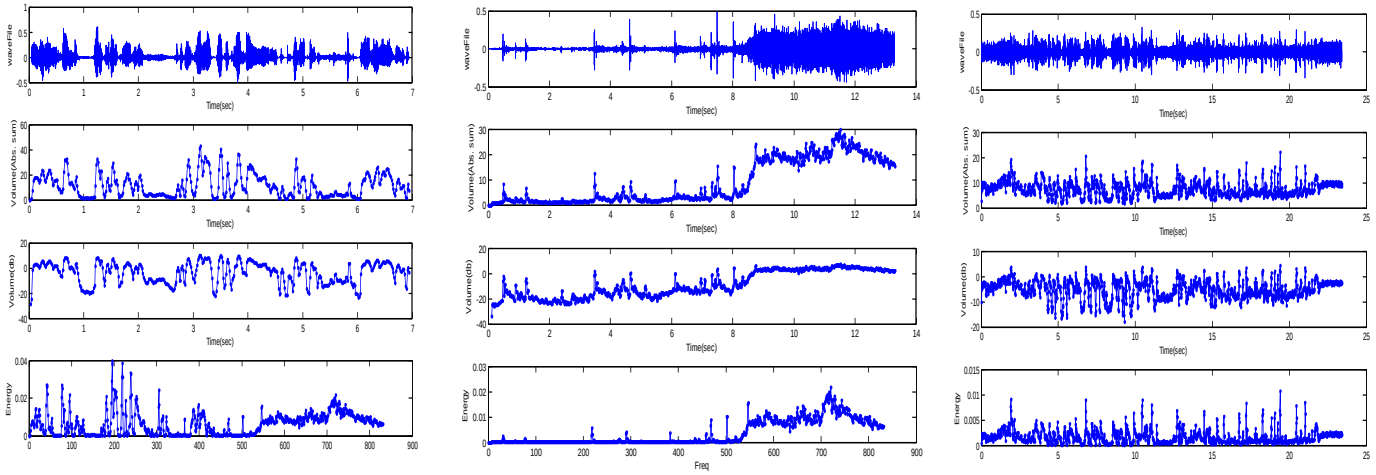


(b) Game 2: Example of PFCS frames and their color histograms



(c) Game 3: Example of PFCS frames and their color histograms

Figure 3. Sample results: Lawn Tennis tournament video Play Field Color Shot frames and their color histograms.

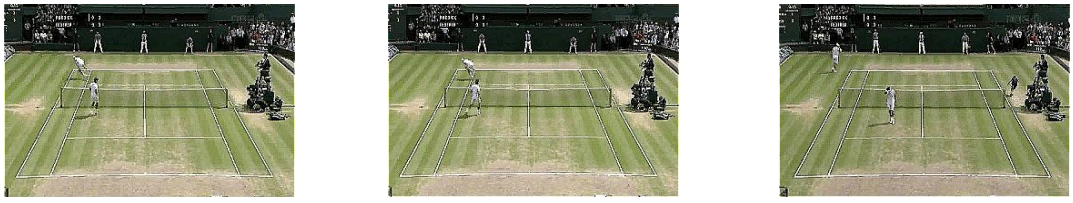


(a) (b) (c)

Figure 4. Examples of lawn tennis audio frame Volume (abs.sum), Volume (Decibel), and Energy of PFCS of French Open, Wimbledon and US Open Lawn

Table 2. Audio Analysis for Video Summarization

Lawn Tennis Tournament	Sample audio shot Number	Time Length of shot (in sec)	Volume (abs. sum)	Volume (Decibel)	Energy	Volume (abs. sum) + Energy
Game1 Wimbledon	01	13	29.9153	7.4859	0.0219	29.9372
Game2 French Open	05	04	45.1370	10.4857	0.0437	45.1807
Game3 US open	05	23	21.9874	4.4110	0.0108	21.9982
Game4 French Open	02	19	24.5997	5.0491	0.0125	24.6122
Game5 Wimbledon	01	11	18.1251	2.9127	0.0076	18.1327



(a) frame # 337 frame # 341 frame # 398

Figure 5. Examples of key frames extracted by our method from lawn tennis tournament video sequence: (a) and (e) Wimbledon, (c) US open, and (b) and (d) French Open tennis.

6. CONCLUSION

In this paper a novel technique for lawn tennis video summarization has been introduced. We developed a shot classification algorithm which is able to efficiently classify shots into PFCS and NPFCS. We analyzed only PFCS and extracted three key frames based on audio, visual and text information from each shot for lawn tennis video summarization. Our experimental results demonstrated that the proposed approach is very efficient and accurate for lawn tennis video.

The future work includes extension of the proposed frame work to different sports such as golf, cricket and table tennis, which require different event detection elements.

7. REFERENCES

[1] Qian Huang, Zhu Liu, Aaron Rosenberg, David Gibbon, and Behzad shahraray, " Automated Generation of News

Content Hierarchy by Integrating Audio, Video, and Text Information," pp. 3025-3028, *IEEE, 1999.*

[2] Rui Cai,Lie Lu,Hong-Jiang Zhang and Lian-Hong Cai, "Highlight Sound Effects Detection in Audio Stream," *ICME 2003,2003 IEEE*,pp.III-37 to III-40.

[3] Xi Shaoo, Changsheng Xu, Mohan S Kankanhalli, "Automatically Generating Summaries for Musical Video," pp.II-547 to II-550, IEEE, 2003.

[4] Baoxin Li, Hao Pan, and Ibrahim Sezan, "A General Framework for Sport Video Summarization with its Application to Soccer," *ICASSP 2003, 2003 IEEE*, pp.III-169 to III-172.

[5] Dian Tjondronegoro and Yi-Ping Phoebe Chen, "Integrating Highlights for More Complete Sport Video Summarization," Published *by the IEEE Computer Society* pp.22-37, *2004 IEEE*.

[6] Chen-Hsiu Huang, Chi-Hao Wu, Jin-Hau Kuo, and Ja-Ling Wu, " A Musical–driven Video Summarization System Using Content-aware Mechanisms," pp. 2711-2714, *IEEE, 2005*.

[7] Rui Cai, Lie Lu, Alan Hanjalic,Hong-Jiang Zhang,Lian-Hong Cai, "A Flexible Framework for Key Audio

Effects Detection and Auditory Context Inference" *IEEE Trans. On Audio, Speech and Language Processing vol.14, No.3*, pp.1026-1039, *May 2006*.

[8] Ahmet Ekin, A. Murat Tekalp, and Rajiv Mehrotra, "Automatic Soccer Video Analysis and Summarization," *IEEE Transactions on Image Processing, vol.12, No.7*, pp.796-807, *July 2003*

[9] N Benjamas, N Cooharojananone, and Chuleerat Jareoskulchai, "Flash light and Player Detection in Fighting Sport for Video Summarization," *Proceedings of ISCIT 2005*, pp. 426-429, IEEE, 2005.

[10] Sandra E.F. de Avila, Antonio da Luz Jr., and Arnaldo de A. Araujo, "VSUMM: A Simple and Efficient Approach for Automatic Video Summarization," IEEE, 2008.

[11] Costas Panagiotakis, Anastasios Doulamis, and Georgios Tziritas, "Equivalent Key Frame Selection Based On Iso-Content Principles," *IEEE Trans. on Circuits and Systems for Video Technology, vol.19, No.3*,pp.447-451,*March 2009*.

[12] Kamesh Namuduri , " Automatic extraction of highlights from a cricket video using MPEG-7 descriptors," IEEE, 2009.

[13] Kenichi Fujimura, Koichiro Honda, and Kuniaki Uehara, "Automatic Video Summarization by using Color and Utterance Information," pp 49-52, IEEE, 2002.

[14] Zhonghua Sun, Kebin Jia, and Hexin Chen, " Video Key Frame Extraction Based on Spatial-temporal Color Distribution," International Conference on Intelligent Information Hiding and Multimedia Signal Processing, , pp 196-199, IEEE, 2008.

[15] Xi Shao, Changsheng Xu, Mohan S Kankanhalli, "A New Approach to Automatic Music Video Summarization," *2004 International Conference on Image Processing (ICIP)*, pp.625-628, *IEEE, 2004*.

[16] Wonjun Kim and Changick Kim, "A New Approach for Overlay Text Detection and Extraction from Complex Video Scene," *IEEE Trans. On Image Processing, vol.18, No.2*, pp.401-411, *February 2009*.

[17] M. R. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," *IEEE Trans.Circuit and Systems for Video Technology*, vol. 15, no. 2, pp. 243–255, Feb. 2005.

[18] Cheolkon Jung, Su Young Lee, and Joongkyu Kim, "Robust Detection of Key Caption for Sports Video Understanding," pp 2520-2523, IEEE, 2008.

[19] Shingo Uchihashi and Jonathan Foote, " Summarizing Video Using A Shot Importance Measure and A Frame Packing-Algorithm", pp 3041-3044, IEEE 1999.

[20] I-Cheng Chang and Kun-You Chen, "Content-Selection Based Video Summarization", IEEE, 2007.

[21] Sudhir S. Kanade, and P. M. Patil, "Representative Frame Decoration Using Unsharp Filter in Video Summarization", ICCSP 2011, IEEE 2011