COVER SHEET

Extraction and Classification of Self-consumable Sport Video Highlights Using Generic HMM

Dian Tjondronegoro [1 2 3], Yi-Ping Phoebe Chen [1], Binh Pham [3]
[1] *School of Information Technology, Deakin University*
[2] *School of Information Systems, Queensland University of Technology*
[3] *Centre for Information Technology Innovations, Queensland University of Technology*
dian@qut.edu.au, phoebe@deakin.edu.au, b.pham@qut.edu.au

Abstract

This paper aims to automatically extract and classify self-consumable sport video highlights. For this purpose, we will emphasize the benefits of using play-break sequences as the effective inputs for HMM-based classifier. HMM is used to model the stochastic pattern of high-level states during specific sport highlights which correspond to the sequence of generic audio-visual measurements extracted from raw video data. This paper uses soccer as the domain study, focusing on the extraction and classification of goal, shot and foul highlights. The experiment work which uses183 play-break sequences from 6 soccer matches will be presented to demonstrate the performance of our proposed scheme.

Keywords: Self-consumable highlights, sport video summarization, Hidden Markov Model (HMM), audio-visual features.

1. Introduction

Sport video highlights extraction and classification has been a challenging topic for researchers around the world during the last decade. The main purpose is to enable effective browsing and retrieval for the rapidly increasing amount of sport video. Most proposed approaches have successfully benefited from the typical structure, recurrent events and specific features of a given sport domain, such as soccer [1] and basketball [2]. However, there is yet a definitive solution to fully automate the selection of 'most appropriate' starting and ending frames for *self-consumable* sport video summary segments. A self-consumable summary scene should include all the required details about an event, such as a goal in a soccer game. In fact, manual intervention seems indispensable to ensure the completeness of a summary scene. To tackle this problem, recent work proposed that *special shot (or frame)* can be used to identify the start of a story, such as 'pitching' view in baseball [3] and 'anchor (news-reader)' view in news videos [4]. Thus, this paper aims to demonstrate that play-break sequences in soccer video (and other similar sports) can be used to identify the boundaries of story units during a game.

After extracting story units, the next step is to classify these units into specific classes, such as goal and foul highlights in soccer video. HMM is a probabilistic model which can be used to model the temporal pattern of a particular highlight [5]. For example, in soccer video, a long global shot (interleaved shortly by other shots) is usually used to describe attacking play which could be paused due to a goal. After a goal is scored, zoom-in and close-up shots will be dominantly used to capture players and supporters celebration. Subsequently, some slow-motion replay shots and artificial texts are usually inserted to add some additional contents to the goal highlight.

Compared to other work on using HMM for highlight classification, this paper uses generic HMM models which are not initialized based on any knowledge of specific sport. Moreover, we use play-break sequences as the inputs for HMM training and classification to achieve consistent length of self-consumable highlight scenes. The measurements include audio-visual features, such as whistle, excitement, near-goal play and text display.

2. Framework

In this paper, we focus on the detection of three types of highlights in a soccer video: *goal*, *foul*, and *shot (goal attempt)*. For this purpose, we identified 5 high-level visual states which are commonly contained by highlights: 1) Global view, 2) Zoom-in view, 3) Close-up view, 4) Near goal, 5) Text Display. In conjunction, we also use 2 audio states: 1) Excitement, 2) Whistle. Although the characteristics of these states vary for different games, they should reveal common statistical similarities of measurements since they are recorded with similar techniques during broadcasting.

When users provide video input with manually-labeled highlight scenes for training, the system will firstly divide each scene into play-break sequences. After all features are extracted from each sequence, the

training algorithm will use the observation sequence to optimize the HMM model which corresponds to the highlight. During classification, the system will use trained models to calculate the most likely highlights (contained by a sequence). Figure 1 illustrates this system architecture.

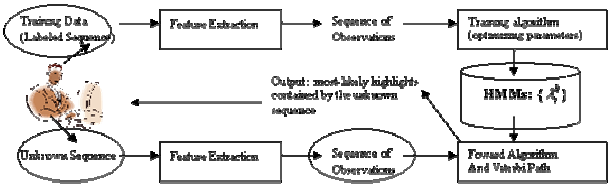

Figure 1. System architecture

3. HMM-based Classifier

To define an HMM model, we need to define the following components:

- Set of (hidden) states, $S_i=\{S_1,S_2...S_N\}$ which corresponds to the remarkable states during highlights. The state at time t is denoted as $q_t \in S$
- Set of measurements (i.e. observations), $M_k=\{M_1,..,M_n\}$ which corresponds to the audio-visual feature sets extractable from video data
- Transition probability matrix between states, $T=\{t_{ab}\}$ where $t_{ab}=P(s_t=s_b \mid s_{t-1}=s_a)$. Constraints: $0 \le t_{ij} \le 1$, and $\sum_{j=1}^{N} t_{ij}=1$
- Measurement probability (or *confusion*) matrix, $C=\{b_j(M_t)\}$ which is the probability of measurement M_k given that the current state is q_t. In a discrete HMM, it is defined $b_j(M_t)=P(M_t=v_k \mid q_t=S_j)$, where $V=\{v_1,v_2...v_K\}$ is the set of all possible observation symbols (thus K is the number of different observation symbols)
- The initial-state transition probability matrix, $\Pi=\{\pi_i\}$, where $\pi_i=P(q_1=s_i)$ which is the probability of a state being the first (i.e. at *t=1*) in a particular model.

In short, an HMM model is represented as a triplet $\lambda=(T,C,\Pi)$.

Training and Classification

For training and classification, a sequence is segmented from play and break shots. For each sequence, we compute audio-visual measurements for each 1-second window. Only the transitions of features are recorded.

Algorithms for initialization and training

- Set the initial models for each highlight (within the scope) as *ergodic* (i.e. each state is connected to all states including itself). There are separate audio and visual models which are denoted as $\lambda_{a_i}^h$ and $\lambda_{v_i}^h$ respectively.
- Set all state transition probabilities, t_{ij} as 1/N where N is the number of states connected to a state. For example if state A is connected to state A, B, C, and D, all transition probabilities would be ¼ (0.25).
- For training, forward-backward algorithm was applied to modify the model parameters which are mostly likely to have generated the observation sequence (i.e. Maximize $P(M_k \mid \lambda)$).

Algorithms for classification and annotation

- For classification, we applied a forward algorithm on each audio and visual model which computes the probability that the observation sequence was produced by the model. This probability is denoted as: α_i^a and α_i^v (thus, we will obtain an array with length equals to the number of highlights within the scope). The combination of audio-visual probability is calculated as $\alpha_i=(\alpha_i^a+\alpha_i^v)/\tau$. In our experiment, τ was set to 3.
- If $\max(\alpha_i) >$ *threshold*, then highlight type h_i is detected, where $i=\arg\max(\alpha_i)$. Otherwise, no highlight is detected (i.e. ordinary play-break).
- Using the specific model which corresponds to the detected highlight, we can use Viterbi path calculation to generate the most likely state sequences for annotation purposes.

Thus, we can use HMM as the connection between high-level concepts and low-level features via the probabilistic matrices. Figure 2 describes this concept.

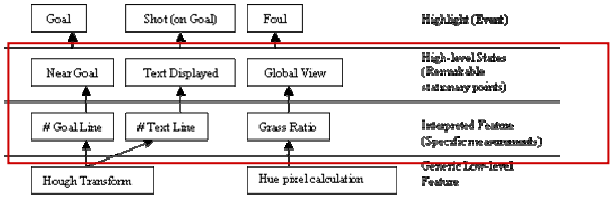

Figure 2. HMM connects high-level concepts to low-level features

4. Features Extraction

In this section, we will describe how play-break sequences are extracted before being classified as

highlights. Subsequently, the features extraction methods used will be summarized.

Play-break Sequence Segmentation

Start of play shot is identified as the first frame of a long global shot (e.g. > 5 sec) which can be interleaved by very short zoom-up or close-up shots (e.g. < 2 sec). Start of break shot is identified as the starting frame of either a long zoom-in shot (e.g. > 5 sec) or a shorter close-up zoom shot (e.g. > 2 sec), which can be interleaved by very short global shots. Moreover, slow-motion replay shots are also regarded as break shots. Slow-motion replay shots are detected by identifying frequent and strong fluctuations in frame difference which is due to the fact that slow motion replay effects are generated by shot-repetition/drop (depending on the camera used during recording) [6]. Using the identified play-start and break start, the system can determine the boundaries of each play and break shots.

A sequence is segmented from the last frame of the last play shot before a break shot until the last frame of the break shot. Figure 3 is provided to clarify this concept. This play-break sequence is chosen as our main input since we noticed that they are effective containers for self-consumable highlights. However, users can extend the play sequence to include more play shots, depending on how much detail on the play they want. Thus, we are reducing the subjectivity level of highlight scope (i.e. compared to the case where users select particular frames). Moreover, instead of retaining the whole play segment, we can improve our data compression by retaining just the play shot which is the closest to break which usually indicate a key event. For viewing purposes, when a goal is detected from a play-break sequence, we propose to retain the first play shot following the last break shot since it usually contains the text display (which gives details about the goal, such as player name and current score line).

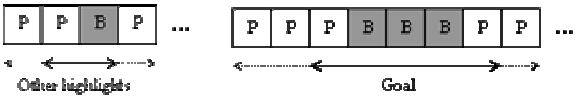

Figure 3. Scope of a play-break sequence

Grass Ratio

Grass (or dominant color)-ratio is a common measurement used for classifying the main shots in soccer video [7]. Global shots contain the highest grass pixels, while zoom-in has less and close-up has the least. To determine the grass-hue index for different videos, we can take random, equally-spread frame samples for an unsupervised training. Since global and zoom-in shots are most dominant, the peak from the total hue-histogram of these random frames will indicate the grass-hue. Grass-hue index must be within 0.15-0.25. This process is repeated 10 times to calculate 10 variations of grass-hue indexes (i.e. $G_1, G_{2,\ldots} G_{10}$). Grass Ratio (GR) is calculated on each frame as:

$$GR = P_G / P$$

where, P_G is the number of pixels which belong to grass-hue and P is the total pixels in a frame. Since there are 10 grass-hue indexes, the final GR is obtained from $\max(GR_1, GR_2, \cdots, GR_{10})$.

Goal Lines

To detect near-goal play, we can detect nearly-parallel lines (as shown in Figure 4) of a certain angle range which are produced by the goal-area lines. These lines are detected as strong peaks in *Hough* transform (compared to a threshold) which is calculated from the gradient image of a frame. Gradient image can be produced either by *Canny* or *Sobel* transform. For soccer, the angle range was set as ℓ and $\Re$, where $100^0 \le \ell \le 120^0$, ℓ is the typical angle range for left-hand-side penalty area, $60^0 \le \Re \le 80^0$, and $\Re$ is the typical angle range for right-hand-side penalty area. The more goal lines we can detect in a frame, the higher probability that the frame shows goal-area.

Text display

In most cases, sport videos use horizontal texts which are surrounded by a rectangular box to emphasize important information. Thus, if we can detect strong horizontal lines which correspond to the text box, we can locate the starting point of a text region. For this purpose, we can re-use the Hough transform on a gradient image. The main difference is that we perform two different checks: the first check ensures that the lines are horizontal (i.e. angle is exactly 90 degrees), the second check is to ensure that the line location is within the usual location for text displays (e.g. less than 90% of the maximum height of the frame).

Whistle Ratio

There are a lot of sports which utilizes whistle to mark specific events. Whistle can be easily identified where there is a strong (spectral) energy inside specific frequency range (e.g. 3500-4500Hz for soccer) which can be calculated as:

$$PSD_W = \sum_{WL}^{WU} | S(n) * conj(S(n)) |$$

where WL and WU is the upper and lower bound of whistle frequency range (respectively), and S(n) is the spectrum (i.e. produced by Fast Fourier Transform) of the audio signal at frequency n Hz. N is the n-point FFT.

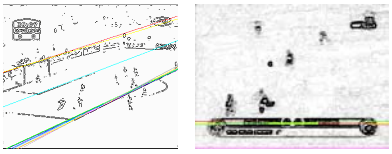

Figure 4. Detectable lines for goal area (left) and text display (right)

Loud-volume, High-pitch and Less-silence Ratio
Loudness, silence and pitch are effective measurements for detecting excitement which are generated by a commentator and/or crowd. We used this equation to calculate the volume of each audio frame: $Volume = \frac{1}{N} * \sum_{n=1}^{N} | x(n) |$

where N is the number of frames in a clip and x(n) is the sample value of the n^{th} frame.

To calculate pitch and silence, we applied the sub-harmonic-to-harmonic ratio based pitch determination in [8] for its reliability. Louder, less silence, and higher pitch audio frames are identified by using dynamic thresholds presented in [9].

5. Experimental Results

We have used a total of 183 play-break sequences from 6 soccer matches from FIFA world cup 2002 and UEFA Champions League 2002-2004. For training, 10 sequences from different game/broadcaster are used for each highlight model. Since the extraction of play-break sequences is essential in our system, Table 1 shows that only a small number of break shots were missed. It is more important to note that Table 1 demonstrates that most highlights can be correctly identified. This performance still can be improved by using more samples for training, as well as improving the accuracy of our features extraction. Details of the features extraction performance can be found in [9].

6. Conclusion and Future Work

We have proposed to use play-break sequences as the effective containers for self-consumable highlight scenes, as well as to achieve more consistent inputs for HMM-based classifier.

DataSet	Play shot	Break shot	Goal	Shot	Important Foul
Juve_Madrid	41 / 42	60 /64	1 /2	8 /10	2 /5
MU_Depor	47 /47	30 /36	2/ 3	4 /7	2 /3
Bra_Ger	24 /25	10 /13	2 /2	2 /3	1 /2
Madrid_Milan	54 /55	35 /40	1 /1	3 /6	1 /1
Milan_Inter	31 /31	36 /37	none	4 /6	1 /2
Arsenal_Loko	43 /45	26 /27	2 /2	4 /7	none

Table1. Performance Results

For future work, we need to extend the experiment using larger datasets (with more variety of structures) for training and classification test. Moreover, users study will be performed to verify that the automatically extracted highlights are self-consumable. Finally, our work will be extended to include more highlights and different sport domains to verify the robustness of our framework.

References

1. Ekin, A. and M. Tekalp, *Automatic Soccer Video Analysis and Summarization.* IEEE Transaction on Image Processing, 2003. **12**(7): p. 796-807.
2. Zhou, W., S. Dao, and C.-C. Jay Kuo, *On-line knowledge- and rule-based video classification system for video indexing and dissemination.* Information Systems, 2002. **27**(8): p. 559-586.
3. Chang, P., M. Han, and Y. Gong. *Extract highlights from baseball game video with hidden Markov models*. in *Image Processing. 2002. Proceedings. 2002 International Conference on*. 2002.
4. Chairsorn, L. and T.-S. Chua. *The Segmentation and Classification of Story Boundaries in News Video*. in *6th IFIP Working Conference on Visual Database Systems*. 2002. Brisbane: Kluwer.
5. Assfalg, J., et al. *Detection and recognition of football highlights using HMM*. in *Electronics, Circuits and Systems, 2002. 9th International Conference on*. 2002.
6. Pan, H., P. van Beek, and M.I. Sezan. *Detection of slow-motion replay segments in sports video for highlights generation*. in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. 2001 IEEE International Conference on*. 2001. Salt Lake City, UT, USA.
7. Xu, P., L. Xie, and S.-F. Chang. *Algorithms and System for Segmentation and Structure Analysis in Soccer Video*. in *IEEE International Conference on Multimedia and Expo*. 1998. Tokyo, Japan,: IEEE.
8. Sun, X. *Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio*. in *ICASSP2002*. 2002. Orlando, Florida.
9. Tjondronegoro, D., Y.-P.P. Chen, and B. Pham. *Sports video summarization using highlights and play-breaks*. in *ACM SIGMM International Workshop on Workshop on Multimedia Information Retrieval (ACM MIR'03)*. 2003. Berkeley, USA: ACM.