Context-Aware Learning For Automatic Sports Highlight Recognition

Bernard Ghanem
King Abdullah University of Science and Technology
bernard.ghanem@kaust.edu.sa

Maya Kreidieh*, Marc Farra*
American University of Beirut
{maf30, mjk10}@ aub.edu.lb

Tianzhu Zhang
Advanced Digital Sciences Centre (ADSC), Singapore
tz.zhang@adsc.com.sg

Abstract

Video highlight recognition is the procedure in which a long video sequence is summarized into a shorter video clip that depicts the most "salient" parts of the sequence. It is an important technique for content delivery systems and search systems which create multimedia content tailored to their users' needs. This paper deals specifically with capturing highlights inherent to sports videos, especially for American football. Our proposed system exploits the multimodal nature of sports videos (i.e. visual, audio, and text cues) to detect the most important segments among them. The optimal combination of these cues is learned in a data-driven fashion using user preferences (expert input) as ground truth. Unlike most highlight recognition systems in the literature that define a highlight to be salient only *in its own right (globally salient), we also consider the context of each video segment w.r.t. the video sequence it belongs to (locally salient). To validate our method, we compile a large dataset of broadcast American football videos, acquire their ground truth highlights, and evaluate the performance of our learning approach.*

1 Introduction

As internet bandwidth increases, the rise of easily accessible multimedia data such as video has created an unprecedented information overflow. Managing video content is different from other media (e.g. text and images) because content providers need to tailor to consumer demands by providing informative previews of the content itself. The objective of video highlight recognition is to condense a relatively long video sequence to a short clip containing the sequence's most salient content. In the sports realm, this would entail a short capture of the game's most salient video segments. The large amount of unprocessed raw sports video holds substantial potential for content-based mining and highlighting applications tendered to sports enthusiasts.

In this paper, we focus on highlights in the sports domain, specifically in the sport of American football. We choose this sport among others due to online availability of videos and highlights, its complex spatiotemporal nature, interest from sports experts (e.g. coaches, trainers, and analysts), and its international appeal. Note that our proposed methodology is easily adaptable to other sports (e.g. tennis and soccer). A video segment's saliency is proportional to its importance at two levels: global and local. Globally, saliency reflects the importance of a video segment with respect to the sport as a whole (e.g. a touchdown is a salient event). In contrast, saliency at the local scale reflects the importance of a video segment with respect to the game (e.g. a last second, game winning score is very salient). Representing a video segment using multi-modal features that depict global and local saliency (e.g. visual, audio, and textual features) is a critical step in measuring the overall saliency of this segment. Moreover, this saliency is usually defined in relation to a particular evaluator. For example, one sports broadcaster may deem salient a play that another broadcaster does not. The function that maps the features describing a video segment to its overall saliency measure is not the same for all evaluators, so it becomes imperative to learn a customizable mapping based on ground truth evaluator preferences.

Related Work

There have been many approaches targeting automatic sports highlight recognition. The most widely explored modalities of highlight creation are audio cues [3, 5], visual cues [6, 7] and text cues [1]. Some researchers combine different modalities, usually by first exploring one type of feature then refining it with others [9, 12]. Game metadata (e.g. game statistics) has been used to denote the importance of a play [10]. Highlights can also be constrained using predefined requirements [2]. Most of these approaches create video highlights by detecting "key events," characterized by predefined attributes such as cheering [9, 4], commentator excitement, or scoring.

Context, with respect to the game and evaluator preferences, is disregarded in most of the previous methods. A game may have no key events, but a descriptive highlight summary is still required. There are recent efforts to incorporate user preferences into the highlight recognition process [1, 12], however, the importance of a video segment is still determined solely by its global saliency, while its context is disregarded.

In this paper, we propose a context-aware learning approach that extracts global and local features from different modalities: audio, video, and text. We propose a data-driven model that learns the relative contribution of each feature in determining overall saliency using evaluator preferences. This measure ranks the plays according to their relevance given the context of the play within the game, and their relevance according to evaluator preferences. The proposed system can thus be easily customized to any evaluator's preferences.

2 Overview of Proposed Method

In this section, we give a brief overview of the components that constitute our highlight system in Figure 1. As input to our system, a sports expert selects a set of videos to preview and selects the relevant highlights within each video. These evaluator preferences are used as ground truth to learn our proposed highlight measure.

2.1 Pre-Processing

We first segment a full-length American football recording into individual plays and remove non-play events (e.g. replays). Each video is accompanied by publicly available text annotations that provide a play-by-play breakdown of the game. We parse these files and extract features relevant to each play (e.g. start time and points scored). To extract meaningful visual and audio features, we map each play to its corresponding sequence of frames in the video. The synchronization between text and video is done automatically by detecting and recognizing the game clock in the video frames.

Detecting the time in each frame can be a computationally expensive task, since American Football games contain a large number of non-play events. However, each broadcaster overlays a banner containing the game clock and other game information, always displayed in the same position within a single broadcast. We locate the game clock by first detecting the banner. Since the banner appears in most of the video frames and remains static, we take a random subset of frames and take the average absolute difference of their grayscale images. After morphological operations, we use adaptive thresholding and connected component analysis to obtain the banner location. To detect which component corresponds to the banner, we learn an SVM classifier using histogram of oriented gradient (HOG) features arranged in a spatial pyramid. The banner classifier reliably discriminates between frames which contain a banner (banner frames) and frames that do not (non-banner frames).

After detecting all banner frames and banner locations, we recognize the game time by using a combination of digit template matching and a Hidden Markov Model (HMM). Given a set of digit templates, we use normalized cross correlation (NCC) to match each digit to a template. We define the NCC template matching score between the i^{th} template $\mathbf{T}_i$ and a clock digit image patch $\mathbf{D}_j$ as x_{ij}. Since these NCC values range between -1 and 1, we define a template-to-digit probability $P(\mathbf{T}_i|\mathbf{D}_j) \propto e^{-\frac{1-x_{ij}}{2}}$. Knowing that the set of possible clock times is simply a subset of the possible digit permutations, the array of probability values is filtered so that it only contains $P(\mathbf{T}_i|\mathbf{D}_j)$ iff $\mathbf{T}_i$ could appear in the j^{th} clock position. The clock time is then inferred from the digit probabilities of each slot in the clock using an HMM. The most probable sequence of templates that fits the underlying image data is efficiently computed using the Viterbi algorithm. The HMM is used to model the spatial transition between two consecutive slots in the clock within the same frame. Temporal transition is also exploited, since the game clock decreases from 15 minutes at the onset of each quarter. Thus, we extend the HMM to model the temporal transition between consecutive game times. Employing an HMM to model the clock's spatiotemporal transitions presents a systematic approach to time detection. It reduces error, is adaptable to any sport after modeling the sport's clock transition pattern, and allows for more robust banner detection by filtering out frames with low game time probability (false positives). Finally, we synchronize the annotated start time of each play (from online text annotation) to its corresponding detected time in the video. As a result, the entire game becomes a sequence of annotated plays.

2.2 Feature Extraction

A total of 82 multimodal features (visual, audio and textual) are extracted from each play to be later combined to determine whether the play is a highlight or not. Each modality provides a complimentary type of information representative of salient events in the video. **(i)** The audio signal of a play is an important indicator of saliency. The commentator's excitement and crowd cheering suggest how salient a particular play is in the context of the entire game. After bandpass filtering to remove noise, a set of widely used time domain and spectral features are then extracted from the audio signal, including the zero crossing rate, short time energy, spectral roll off, spectral centroid/flux, and MFCC coefficients. **(ii)**. Play-by-play textual annotations hold valuable semantic information. They are parsed for all relevant information and stored in a database. With prior knowledge of the game hierarchy, this textual informa-

Figure 1. Flowchart of the proposed highlight recognition system

tion can be used to look at particular types of plays (e.g. passes, touchdowns, etc.). These features play an important role in determining the global saliency of a play. Also, play statistics (e.g. where the play occurred and how many points were scored) can be extracted to determine the local saliency of each play. **(iii)** Since text cues already annotate the semantic content of the play themselves (e.g. the main events that occur within a play), we refrain from building computationally expensive automatic activity detectors. Instead, we use visual data to determine whether a play is followed by a replay. Commentators usually initiate a replay immediately after an interesting play. As such, it is a visual mechanism to convey saliency to viewers. We use the previous banner detection results to recognize non-play segments containing replays (i.e. non-banner frames between two plays). The length of a replay is indicative of the saliency of the preceding play.

2.3 Learning Highlights

Given a set of N training videos and its corresponding set of highlight plays, we seek to learn a parametric mapping that predicts the measure of a play being a highlight within a test video V_T. Each video sequence may contain a different number of plays, each of which is represented by an 82 dimensional feature vector as described earlier. Each play in the training set is labeled either as $+1$ indicating a highlight or -1 indicating a non-highlight. Previous learning-based highlight recognition methods naively pool all plays together and learn a classifier that discriminates between highlight and non-highlight plays irrespective of the videos they belong to (i.e. irrespective of their context). As mentioned earlier, a play's saliency (highlight measure) is highly dependent on its context. For example, two plays with exactly the same features may be labeled differently, since they belong to different contexts (game videos). In other words, a play may be considered a highlight in one game but a non-highlight in another. Consequently, we argue that building a context-aware highlight classifier provides more discriminative power than a naive classifier that is oblivious of context. As such, we believe that the highlight measure of V_T should be influenced more by training videos that share similar play content.

In this paper, we define a similarity measure between two game videos, which estimates play content similarity. This measure is based on the Fisher kernel model for representing groups of feature vectors. Due to its excellent performance in computer vision tasks [8], we choose the Fisher kernel paradigm over other models (e.g. Bag-Of-Words). To do this, we build a *universal*, parametric probability distribution for *all* plays in the training set. In our experiments, we choose this distribution to be a Gaussian Mixture Model (GMM), which can be learned using a conventional EM algorithm. Each play in the training set is defined by a Fisher kernel vector, which is the gradient of the universal GMM (w.r.t. its parameters) evaluated at the play's feature vector. As such, each game video is represented as a linear combination of the Fisher kernel vectors corresponding to its constituent plays. The similarity measure between two games is computed as the similarity (normalized correlation) between their Fisher kernel representations.

Then, we build a set of highlight classifiers (kernel SVM), each corresponding to a cluster of *similar* videos in the training set. For simplicity, we take each cluster to be a singleton. To classify each play $\vec{\mathbf{p}}_j$ in an unseen video V_T, we build a customized classifier C_T for V_T. This classification scheme is similar in spirit to the SVM-kNN method proposed in [11] for image classification. The proposed classifier is defined as a convex combination of the training classifiers each weighted by the similarity of V_T to the classifier's corresponding set of training videos (refer to Figure 2). Finally, each play $\vec{\mathbf{p}}_j$ in V_T is assigned a highlight measure or saliency score $C_T\left(\vec{\mathbf{p}}_j\right)$. Then, a predefined threshold θ (usually $\theta = 0$) is applied to this measure to determine whether $\vec{\mathbf{p}}_j$ is a highlight or not. To produce a context-aware highlight measure for all plays within the same game, we exponentiate the previous measures and normalize all play measures to sum to 1. This also generates the probability of each play being a highlight in the context of V_T. We can threshold this probability at different levels to generate different sized highlight summaries.

3 Experiments

In this section, we validate our highlight recognition system in the realm of American football. For this purpose, we compiled a large scale dataset of broadcast American football videos from online sources. This dataset comprises of 150 games roughly 2.5 hours each in length. Each game has been automatically synchronized with its ground truth text annotation and segmented into its constituent plays. Ground truth play

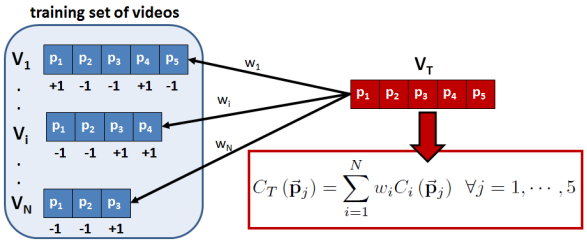

Figure 2. Context-aware highlight recognition

highlights were compiled from online expert sources (e.g. ESPN and NFL), which depict the important plays in each game. Each play is represented using the 82 features described above. As another contribution, we seek to make these features and the ground truth publicly available. This allows for objective quantitative evaluation/comparison of highlight methods.

The cascade approach of detecting the banner then detecting the game clock results in a more efficient and robust system. In fact, this approach results in a significant spatial and temporal reduction of the search space required for time detection. Detecting the banner and removing the non-banner frames reduces the number of video frames to be processed by 30% and the number of pixels by 90%, resulting in processing runtime of 15 sec/min of video for time detection.

We use the classification scheme outlined in Section 2.3 to learn the overall highlight measure. To validate the importance of context in highlight recognition, we compare our context-aware classifier to the naive baseline, which learns a global highlight (SVM) classifier after pooling all plays in the training set and discarding context information (refer to Section 2.3). Half the dataset is used for training, with the rest for testing. This division is randomized 10 times and average classification performance is recorded. In Figure 3, we plot the average ROC results for the proposed and baseline methods. This plot is obtained by varying the classification threshold θ. Clearly, our method exploits context to outperform the baseline classifier. In fact, since highlight plays are comparatively quite sparse (about 7%) in the dataset, the behavior/performance of highlight classifiers at high false positive rates is important. As the false position rate increases (i.e. more non-highlight plays are allowed to be classified incorrectly), it is desirable that the true positive rate increases as quickly as possible (i.e. the derivative of the ROC curve is high). This is more the case with our proposed classifier than the baseline. Moreover, the low overall accuracy of both methods points out the difficulty of this classification problem and the potential for future improvement.

4 Conclusion

This paper proposes a method to automatically generate highlights for broadcast American football games.

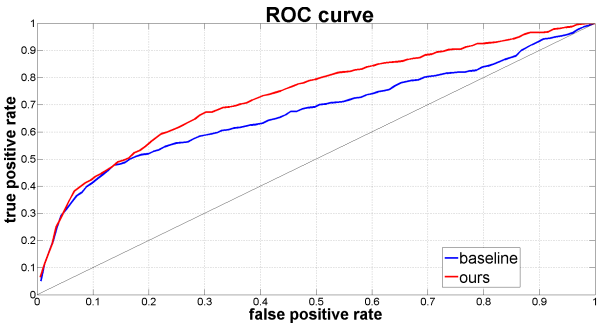

Figure 3. ROC results for the proposed and baseline highlight recognition methods

Each game segment is represented by multimodal (visual, audio, and text) features that are indicators of its overall (global and local) saliency. Using a context-aware boosted learning technique, we learn how these features should be combined to discriminate highlight plays from non-highlights within their respective contexts. Our experiments on a large dataset of real-world videos show significant performance over the baseline.

Acknowledgments: This study is supported by the research grant for the Human Sixth Sense Program at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).

References

[1] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi. Personalized Abstraction of Broadcasted American Football Video by Highlight Selection. *IEEE Transactions on Multimedia*, 6(4):575–586, Aug. 2004.

[2] M. Barbieri, N. Dimitrova, and L. Agnihotri. Movie-in-a-Minute : Automatically Generated Video Previews. *Intelligent Algorithms in Ambient and Biomedical Computing*, pages 89–101, 2004.

[3] H. Duxans, X. Anguera, and D. Conejero. Audio based soccer game summarization. *2009 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pages 1–6, May 2009.

[4] A. Hanjalic. Adaptive extraction of highlights from a sport video based on excitement modeling. *IEEE Transactions on Multimedia*, pages 1114–1122, 2005.

[5] H. Harb and L. Chen. Highlights detection in sports videos based on audio analysis. In *CBMI*, 2003.

[6] Y. S. Kumar, S. K. Gupta, B. R. Kiran, K. R. Ramakrishnan, and C. Bhattacharyya. Automatic summarization of broadcast cricket videos. In *International Symposium on Consumer Electronics*, volume 3, pages 222–225, 2011.

[7] G.-G. Lee, H.-k. Kim, and W.-Y. Kim. Highlight generation for basketball video using probabilistic excitement. In *International Conference on Multimedia and Expo*, pages 318–321, June 2009.

[8] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. In *ECCV*, pages 143–156. Springer, 2010.

[9] Y. Song and W. Wang. Unified Sports Video Highlight Detection Based on Multi-feature Fusion. *International Conference on Multimedia and Ubiquitous Engineering*, pages 83–87, June 2009.

[10] Y. Takahashi, N. Nitta, and N. Babaguchi. Automatic Video Summarization of Sports Videos Using Metadata. *Language*, pages 272–280, 2004.

[11] H. Zhang, A. Berg, and M. Maire. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. *CVPR*, 2006.

[12] Y.-F. Zhang, C. Xu, X. Zhang, and H. Lu. Personalized retrieval of sports video based on multi-modal analysis and user preference acquisition. *Multimedia Tools and Applications*, 44:305–330, 2009.