The Excitement of Sports: Automatic Highlights Using Audio/Visual Cues

Michele Merler[1] Dhiraj Joshi[1] Khoi-Nguyen C. Mac[2] Quoc-Bao Nguyen[1] John Kent[3]
Stephen Hammer[3] Jinjun Xiong[1] Minh N. Do [2] John R. Smith[1] Rogerio S. Feris[1]

[1]IBM T. J. Watson Research Center [2] University of Illinois at Urbana-Champaign [3] IBM iX

Abstract

The production of sports highlight packages summarizing a game's most exciting moments is an essential task for broadcast media. Yet, it requires labor-intensive video editing. We propose a novel approach for auto-curating sports highlights, and demonstrate it to create a first of a kind, real-world system for the editorial aid of golf and tennis highlight reels. Our method fuses information from the players' reactions (action recognition such as high-fives and fist pumps), players' expressions (aggressive, tense, smiling and neutral), spectators (crowd cheering), commentator (tone of the voice and word analysis) and game analytics to determine the most interesting moments of a game.

1. Introduction

The tremendous growth of video data has resulted in a significant demand for tools that can accelerate and simplify the production of sports highlight packages for more effective browsing, searching, and content summarization. We present a novel approach for auto-curating sports highlights, showcasing its application to major golf and tennis tournaments (Masters, Wimbledon and US Open). Our approach combines information from the *player*, *spectators*, *commentator* and *game analysis* to determine a game's most exciting moments. Video segments are then added to an interactive dashboard for quick review and retrieval by a video editor or broadcast producer, speeding up the process by which these highlights can then be shared with fans. Figure 1 shows the interface of our system, called High-Five (**High**lights **F**rom **I**ntelligent **V**ideo **E**ngine), H5 in short.

The first prototype of IBM H5 [5, 3] was deployed at the 2017 Masters golf tournament extracting highlights live from multiple video streams over four days. Based on its success, H5 was further adapted to tennis content to be employed during the 2017 Wimbledon and US Open tennis tournaments. Personalized highlight extraction and retrieval is another unique feature of our system, using meta-data information about players extracted automatically from video graphics or provided by analysts and court-side statisticians

Several methods have been proposed to automatically extract highlights from sports videos based on audio and visual cues, with approaches using the analysis of replays, crowd cheering, closed captioning and social media reactions [6, 8, 7]. More recently, Bettadapura et al. [2] used contextual cues from the environment to understand the excitement levels within a basketball game. Our proposed approach offers a unique combination of excitement measures to produce highlights, including information from the *spectators*, the *commentator*, and the *player* reaction.

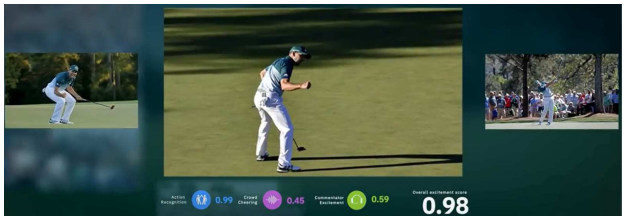


Figure 1. The H5 system dashboard for auto-curation of sports highlights. Highlights are identified in near real-time with an associated excitement level score. The user can click on the icons in the panel to play the associated video in the center, along with the scores for different excitement measures.

2. Excitement Markers

2.1. Audio and Text based Markers Detection

Crowd cheering is perhaps the purest form of approval of a player's shot within the context of any sport. Another important one is excitement in the commentators' tone while describing a shot. Together those two audio cues play a key role in determining the position and excitement level of a potential highlight clip. We leverage SoundNet[1] to construct audio-based classifiers for crowd-cheering and commentator tone excitement. It uses a deep 1-D convolutional neural network architecture to learn representations of environmental sounds from nearly 2 million unlabeled videos. We learn linear SVM models atop the deep features to classify crowd cheer/commentator tone excitement. The tone based commentator excitement measure is complemented by a text-based excitement marker which leverages a text-to-speech engine and uses a dictionary of 60 expressions indicative of excitement (e.g. "great shot", "fantastic").

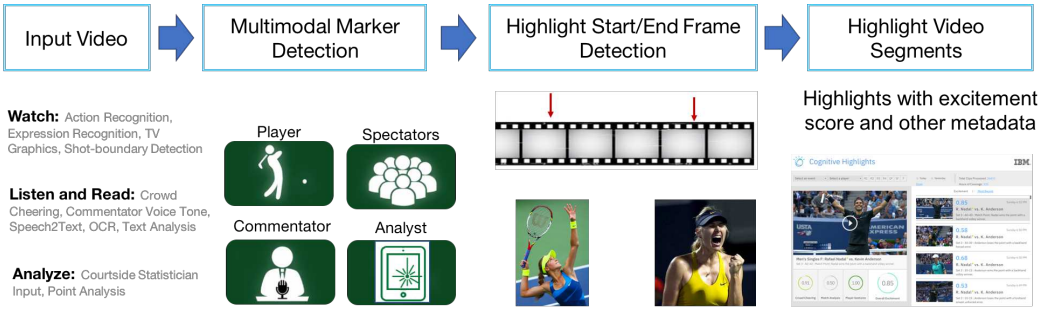


Figure 2. Our approach consists of applying multimodal (video, audio, text) marker detectors to measure the excitement levels of the player, spectators, and commentator in video segment proposals. The start/end frames of key shot highlights are accurately identified based on these markers, along with the detection of TV graphics (when available as in golf) and visual shot boundaries, or information from court-side statisticians. The output highlight segments are associated with an overall excitement score as well as additional metadata about the video segment such as the player name, hole number and shot information in golf, or match point information in tennis.

2.2. Visual Marker Detection

Player Reaction is another important cue to determine interesting moments of a game. In our work, we train an action recognizer to detect a player celebrating. Inspired by [4], we use still images which are much easier to annotate and allow training with less computational resources compared to video-based classifiers. At test time, the classifier is applied at every frame and the scores aggregated for the highlight segment. Classifiers to detect player's celebration are based upon the VGG-16 and the ResNet-50 architectures pretrained on Imagenet. Positive examples are sampled from 2016 Masters, Wimbledon, and US Open videos, and also from the web. Negative examples are sampled from the videos, as explained in Section 2.4.

Facial Expression carries valuable information that can augment or correct predictions from the player reaction models. For example, a tennis player might be raising his arm to collect a ball instead of celebrating a point. In this case, detecting a neutral facial expression can help rejecting a false positive instance. Training data to build a facial expression classifier was collected by extracting faces from the action celebration training images, using a SSD detector. The extracted faces were then categorized into four types of expression: aggressive, tense, smiling, and neutral.The first three associated with celebration, the last considered as non-celebratory. The classifier was trained by fine-tuning a VGG-face model on a dataset of tennis players faces.

2.3. Game Analytics

In tennis not every point has equal relevance within a game. For example *match points* and *set points* are more valuable than others, and business rules require them to be included in official highlights packages. During the tournaments we received live information about the points from side court statisticians and compiled it into a single analytics score in the following manner, which was devised following expert advice concerning the significance and difficulty of each item: **(1)** -0.1 for a point won due to unforced error or rally count smaller than 3, **(2)** +0.1 for a point won due to positive play, volley winner, smash winner, match point, break point, or rally count greater than 5, **(3)** +0.20 for a point won due to forced error, player movement detected, or rally count greater than 10, **(4)** +0.25 for a game winning point. Positive play means a point won thanks to a player's active effort, not an opponent's mistake.The sum of values for any given point was then normalized in the range 0 to 1.

2.4. Cross Modality Bootstrapping

While the audio and visual classifiers were independent modules within the system, the training data gathering process proceeded through several rounds of bootstrapping, which exploited the correlation among modalities. In particular we applied this principle for the visual player celebration classifier training. The underlying assumption is that the likelihood of finding frames showing players celebrating is higher in frames close to a detected crowd cheer. As shown in Figure 3 in each bootstrapping round, we fed unlabeled videos to the current classifiers, and sent to annotate the samples with highest positive or negative scores for each individual classifier, plus for the visual one the samples close in time with the highest scores of the cheer model. The newly labeled data was than used to finetune the classifiers for the next bootstrapping round, until a certain accuracy level on a held out validation set was reached.

3. Experiments

3.1. Experimental Setting

We evaluated our system in three real world championships, namely the 2017 Masters, Wimbledon, and US Open tournaments. For Masters, we analyzed in near real-time the content of four channels broadcasting simultaneously over the course of four consecutive days, for a total of 124 hours of content. The system ran on a Redhat Linux box with two K40 GPUs and produced 741 highlights in total. We extracted frames directly from the video stream at a rate of 1fps and audio in 6 seconds segments encoded as

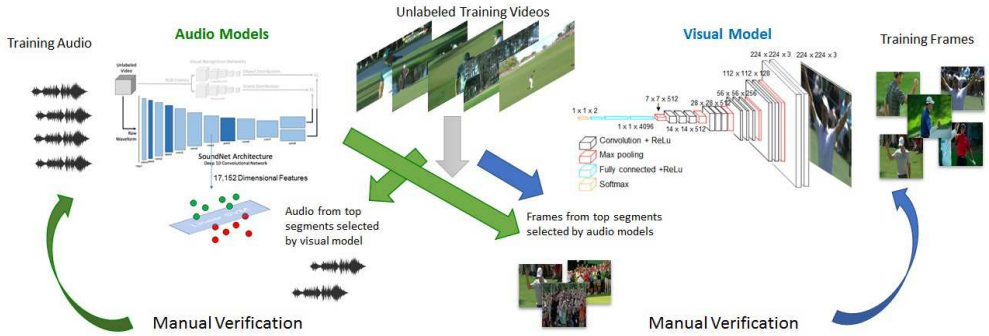


Figure 3. Bootstrapping loop of data curation and fine-tuning of the individual markers models. Data for each bootstrapping round is selected based on the scores of the previous round and other modalities models (from audio for the player celebration).

16bit PCM at rate 22,050. The Wimbledon and US Open system ran on two Ubuntu nodes with four K80 GPUs each providing a total of 16 stream services to rank candidate highlight clips during the tournaments. Videos were chunked in 10 seconds segments and analyzed in less than 2.5 seconds each through our service APIs.

3.2. Individual Markers

All training for the individual markers was performed on content from the 2016 tournaments videos and images downloaded from the web, while testing was done on video data from the 2017 tournaments.

Player Celebration: We used Caffe for training VGG-16 and ResNet models with stochastic gradient descent. The data collected via multiple rounds of bootstrapping was augmented by random cropping and horizontal flipping. For Masters we used 2,906 positive examples and 6,744 negative ones, for Wimbledon 13,263 and 33,372, and for US Open 11,330 and 12,516. We evaluated on clips randomly selected from each of the 2017 tournaments and manually labeled. The number of frames, positive examples and negative examples for each tournament were 1,064, 59 and 1,005 for Masters, 4,777, 78 and 4,699 for Wimbledon, 8,963 and 52 and 8,911 for US Open, respectively. The imbalance of positive and negative examples reflects the actual distribution of data, since occurrences of a player celebrating are relatively rare within a match. Classification accuracies on 2017 Masters, Wimbledon, and US Open data were 98.4%, 98.12% and 99.33% respectively, while AUC were 0.9018, 0.9465, and 0.9313.

Facial expression: This marker was tested on faces extracted from frames in the 2017 tennis test set videos. Expressions on the players faces were manually labeled as *excited* (combining aggressive, tense and smiling) or *neutral*. For Wimbledon we tested on 112 excited and 360 neutral faces, while for US Open 230 and 889. The models performed reliably for both tournaments, with Wimbledon's performance being better (AUC of 0.81 versus 0.75). Recognition accuracies of expressive against neutral expression were 82.42% and 82.23%, respectively.

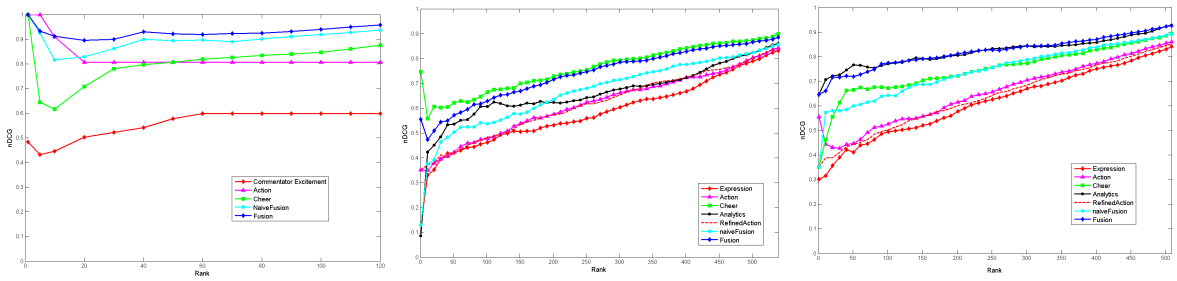
Crowd Cheering Marker: Cheer samples from 2016 Masters and Wimbledon replay videos as well as examples of cheer obtained from YouTube were used in order to train the audio cheer classifier using a linear SVM on top of deep features. For negative examples, we used audio tracks containing regular speech, music, and other non-cheer sounds found in Masters and Wimbledon replays. In total our final training set consisted of 453 positive and 454 negative samples (6 seconds each). We manually annotated random sets of six-seconds audio clips from the 2017 Masters (69 cheer, 336 non-cheer), Wimbledon (158, 915) and US Open (627, 937) tournaments videos to evaluate the performance of the model. The performance of the audio cheer model was measured as AUC of 0.9, 0.94 and 0.93 for 2017 Masters, Wimbledon and US Open, respectively.

Commentator Tone Marker: Similarly to crowd cheering, we created a training set for the commentator tone excitement marker using 2016 videos and several rounds of bootstrapping.The final training set consisted of 131 positive and 217 negative samples. We evaluated on a set of 240 randomly sampled audio snippets from the 2017 Masters and manually labeled as excited tone or not, and the model obtained AUC = 0.77. While the performance of this model is not as good as the cheer classifier, it was reliable enough to be employed in the live system during the tournament.

3.3. Highlights Detection

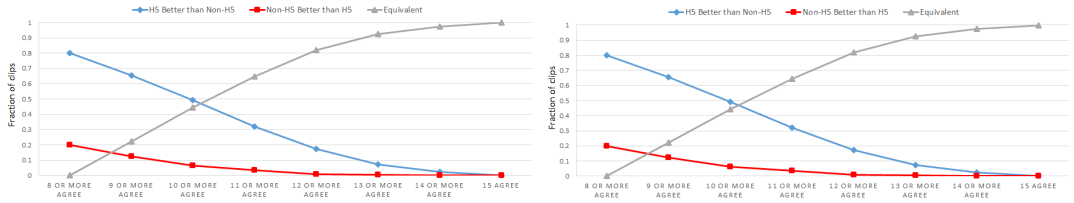
Evaluating the quality of sports highlights is a challenging task, since a clearly defined ground truth does not exist. We approached this problem by comparing the clips automatically generated by our system to human judgments [2].

Human Rankings: We conducted user studies on Amazon Mechanical Turk in which workers were asked to evaluate the excitement level of several clips randomly sampled from the ones selected and scored by H5. We asked each participant to assign a score from 0 to 5 to a clip, with 0 meaning no interesting content and 5 being the most exciting shots. We then averaged the scores of the users for each clip. The resulting scores determined that 92.68% of the clips produced by our system were legitimate high-



(a) 2017 Masters (b) 2017 Wimbledon (c) 2017 US Open

Figure 4. nDCG computed at different ranks for the individual components as well as the Fusion.



(a) 2017 Wimbledon Preferences (b) 2017 US Open Preferences

Figure 5. Human preferences in AB tests for 2017 Wimbledon and US Open.

lights (scores 2 and above), while 7.32% were mistakes. We then compared the rankings of the clips according to the scores of each individual marker, as well as their fusion, to the ranking obtained through the users votes. The performance of each ranking is computed at different depth k with the normalized discounted cumulative gain (nDCG) metric (Figure 4). Naive-Fusion used equal weights to combine the normalized scores from each component, while Fusion (used by the live system) used weights optimized through cross-validation on a separate training set. For all tournaments the system's Fusion outperforms the Naive one.

A/B Testing: Besides the ranking, for Tennis we also wanted to determine whether the *selection* made by the system about which clips should go into the compiled highlights aligned with human preferences. Thus we evaluated the clip selection process through another AMT experiment. For each tournament we randomly selected 500 pairs of clips. In each pair both clips belonged to the same game: one clip which had been selected to be part of the highlights, and one clip which had been discarded. We presented each pair to the workers and asked them to pick which clip in the pair was more exciting and/or interesting. Each pair was voted on by 15 workers, and a total of 234 unique users participated in the study. From Figure 5 (a) and (b) we can observe how for both tournaments the majority of voters picked the clips which were selected by the system to be part of the highlights of a game (blue curves) overwhelmingly over the non-highlight worthy ones (red curves). Naturally the fraction of clips on which a larger number of users agrees decreases as we move from 8 (the majority of voters) to to 15 (all the voters), a trend clearly visible in the growth of the grey curves representing an indecision.

4. Conclusion

We presented a novel approach for automatically extracting highlights from sports videos based on multimodal sport-independent excitement measures, for which models were learned with reduced cost in training data annotation by exploiting the correlation of different modalities. We demonstrated the first-of-a-kind H5 system in three major golf and tennis tournaments in 2017, and showed that it agrees with human preferences.

References

[1] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations in unlabeled video. In *NIPS*, 2016.

[2] V. Bettadapura, C. Pantofaru, and I. Essa. Leveraging contextual cues for generating basketball highlights. In *ACM Multimedia*, 2016.

[3] D. Joshi, M. Merler, Q.-B. Nguyen, S. Hammer, J. Kent, J. Smith, and R. Feris. Ibm high-five: Highlights from intelligent video engine. In *ACM Multimedia*, 2017.

[4] S. Ma, S. A. Bargal, J. Zhang, L. Sigal, and S. Sclaroff. Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern Recognition*, 2017.

[5] M. Merler, D. Joshi, Q. B. Nguyen, S. Hammer, J. Kent, J. R. Smith, and R. S. Feris. Automatic curation of golf highlights using multimodal excitement features. In *CVPRW*, 2017.

[6] A. Tang and S. Boring. # epicplay: Crowd-sourcing sports video highlights. In *ACM CHI*, 2012.

[7] D. Zhang and S.-F. Chang. Event detection in baseball video using superimposed caption recognition. In *ACM Multimedia*, 2002.

[8] Z. Zhao, S. Jiang, Q. Huang, and G. Zhu. Highlight summarization in sports video based on replay detection. In *ICME*, 2006.