Automatic Curation of Sports Highlights using Multimodal Excitement Features

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

***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. We accurately identify the start and end frames of key shot highlights with additional metadata, such as the player's name and the hole number, or analysts input allowing personalized content summarization and retrieval. In addition, we introduce new techniques for learning our classifiers with reduced manual training data annotation by exploiting the correlation of different modalities. Our work has been demonstrated at a major golf tournament (2017 Masters) and two major international tennis tournaments (2017 Wimbledon and US Open), successfully extracting highlights through the course of the sporting events [1–3]. 54% of the clips selected by our system overlapped with the official highlights reels for the 2017 Masters. Furthermore, user studies showed that people found that 90% of the non-overlapping ones where of the same quality of the official clips for 2017 Masters, while the automatic selection of clips for highlights of 2017 Wimbledon and 2017 US Open agreed with human preferences 80% and 84.2% of the time, respectively.**

I. 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. In a major professional golf tournament such as Masters, for example, with 90 golfers playing multiple rounds over four days, video from every tee, every hole and multiple camera angles can quickly add up to hundreds of hours of footage. Wimbledon, the oldest tennis tournament, hosts around 250 singles matches alone over the course of 13 days again producing several hundreds of hours of video. Yet, most of the process for producing highlight reels in those tournaments is still manual, labor-intensive, and not scalable.

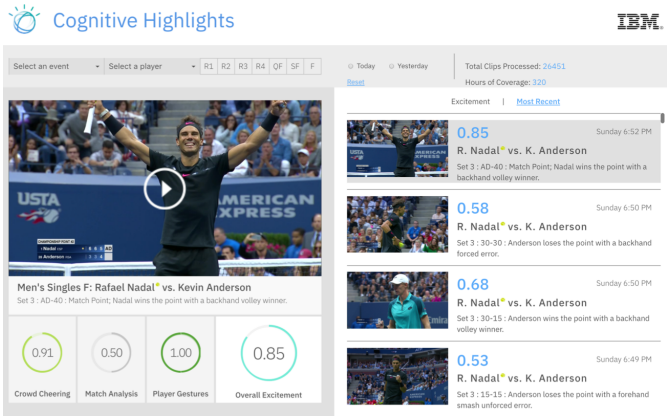


Fig. 1. The H5 system dashboard for auto-curation of sports highlights. Highlights are identified in near real-time (shown in the right panel) with an associated excitement level score. The user can click on the icons in the right panel to play the associated video in the center, along with the scores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

In this paper, we present a novel approach for auto-curating sports highlights, showcasing its application for golf (2017 Masters) and tennis (2017 Wimbledon and US Open). Our approach combines information from the *player*, *spectators*, and the *commentator* to determine a game's most exciting moments. We measure the excitement level of video segments based on the following main multimodal markers:

- **Player reaction:** visual action recognition of player's celebration (such as high fives or fist pumps) and facial expression recognition;
- **Spectators:** audio measurement of crowd cheers;

Manuscript submitted for review on 02/08/18

Michele Merler (corresponding author), Dhiraj Joshi, Quoc-Bao Nguyen, Jinjun Xiong, John R. Smith and Rogerio S. Feris are with IBM Research, Yorktown Heights, New York, NY, 10598, USA. E-mail: {mimerler,djoshi,quocbao,jinjun,jsmith,rsferis}@us.ibm.com

Khoi-Nguyen C. Mac and Minh N. Do are with the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. E-mail: {knmac,minhdo}@illinois.edu

Stephen Hammer and John Kent are with IBM iX, New York, NY 10003, USA. E-mail: {johnkent,hammers}@us.ibm.edu

- **Commentator:** excitement measure based on the commentator's tone of the voice, as well as exciting words or expressions used, such as "beautiful shot".

For golf, these indicators are used along with the detection of TV graphics (e.g., lower third banners) and shot-boundary detection to accurately identify the start and end frames of key shot highlights with an overall excitement score. For tennis, the start and end points for an exemplar highlight shot can be accurately determined based on input from court-side statisticians and analysts who actively annotate tennis matches in real time. 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 those highlights can then be shared with fans eager to see the latest action. 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 [15, 20] was deployed at the 2017 Masters golf tournament, extracting highlights live from multiple video streams over the course of four days. Based on its success, H5 was further adapted to tennis content and employed during the 2017 Wimbledon and US Open tennis tournaments. This adapted H5 prototype introduced the use of *player expression:* expression on the face of the tennis player (i.e. aggressive, tense, smiling, neutral) to improve the player's reaction marker. Based on the observation that tennis commentary tends to be less colorful, quite to the point, and rarely excited in tone, the tennis H5 prototype did not employ commentator based markers. The system was successfully employed as the official highlights provider for the Wimbledon and US Open tennis tournaments in 2017.

Besides incorporating multimodal (audio, visual, text) and multi-source (crowd audio, commentator speech, player body, player face, overlaid text, speech text) information for highlights detection, we also exploit how one modality can guide the learning of another modality, with the goal of reducing the cost of manual training data annotation. In particular, we show that we can use TV graphics and OCR as a proxy to build rich feature representations for golf player recognition from *unlabeled* videos, without requiring costly training data annotation. Our audio-based classifiers also rely on feature representations learned from unlabeled video [5], and are used to constrain the training data collection of other modalities (e.g., we use the crowd cheer detector to select training data for player reaction recognition).

Personalized highlight extraction and retrieval is another unique feature of our system. In golf, by leveraging TV graphics and OCR, our method automatically gathers information about the golf player's name and the hole number. This metadata is matched with relevant highlight segments, enabling searches like "show me all highlights of player X at hole Y during the tournament" and personalized highlights generation based on a viewer's favorite players. For tennis, information about players is extracted by meta-data provided by analysts and court-side statisticians, thus allowing the same type of personalization when combined with the analyzed video.

The key **contributions** of our work are listed below:

- We present a first-of-kind system for automatically extracting sport highlights by uniquely fusing multimodal excitement measures from the player, spectators, and commentator. In addition, by either automatically extracting metadata via TV graphics and OCR or obtaining it from court-side statisticians, we allow personalized highlight retrieval or alerts based on player name, hole or field number, location, and time.
- We introduce novel techniques for learning our multimodal classifiers without requiring costly manual training data annotation. In particular, we build rich feature representations for player recognition without manually annotated training examples.
- We provide an extensive evaluation of our work, showing the importance of each component in our multimodal approach through ablation studies. We compare our results with professionally curated golf highlights. We also provide an extensive user study comparing highlights automatically produced of our H5 system to human preferences by employing Amazon Mechanical Turk, for both golf and tennis.

Our system was successfully demonstrated and deployed at major international golf and tennis tournaments in 2017, extracting highlights from multiple live channels during the course of the tournaments [1–3].

II. Related Work

Video Summarization. There is a long history of research on video summarization [18, 25, 42], which aims at producing short videos or keyframes that summarize the main content of long full-length videos, by looking at elminating redundancy either at signal level (feature dimensionality reduction [41]) or in semantic content [42]. Our work also aims at summarizing video content, but instead of optimizing for representativeness and diversity, as traditional video summarization methods do, our goal is to find highlights or exciting moments in the videos. A few recent methods address the problem of highlight detection in consumer videos [31, 38, 39]. Instead our focus is on sports videos,

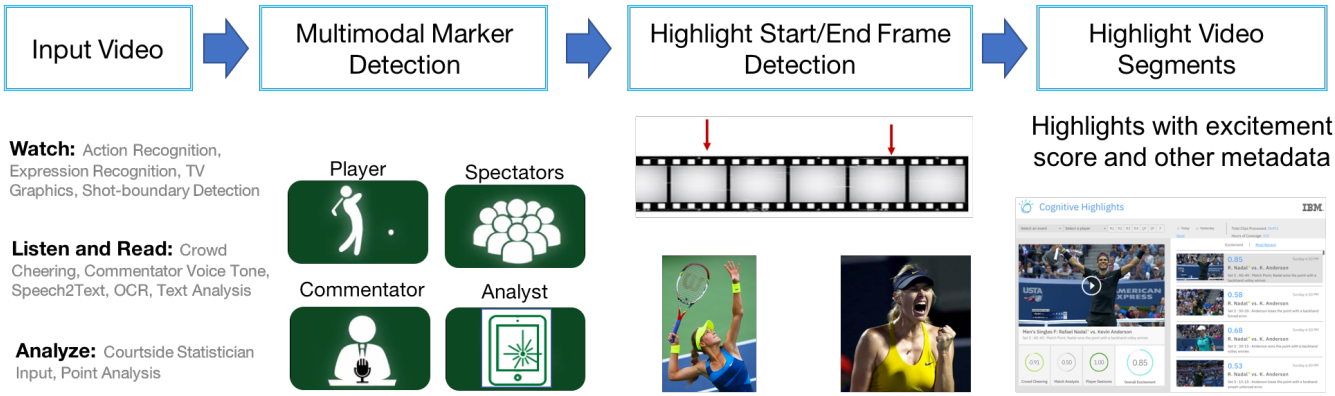


Fig. 2. Our approach consists of applying multimodal (video, audio, text) marker detectors to measure the excitement levels of the player, spectators, and commentator in video segment proposals. The start/end frames of key shot highlights are accurately identified based on these markers, along with the 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.

which offer more structure and objective metrics than unconstrained consumer videos.

Automatic Trailer Generation. Another sub-area of video summarization involving multimodal video analysis that goes beyond content recognition, and focusing instead on affective responses evoked by the video, is movie trailer generation [9, 11, 37]. For example, Evangelopoulos et al. [9] model and combine audio, visual and textual saliency to select the most relevant scenes in a movie. In this space, works focus on detecting content with the highest emotional impact based on movie genre. For instance, in horror movies scenes evoking feelings of suspense or fear are important [29]. In our domain of interest, on the other hand, only positive emotions connected to excitement are relevant. Furthermore, differently from this line of research, the focus of our work is on identifying and measuring subjects reactions (players, crowd, and commentator) directly in the video stream, rather than inferring reactions which are supposed to be evoked by inspected content which is deemed as "impressive" [11].

Sports Highlights Generation. Several methods have been proposed to automatically extract highlights from sports videos based on audio and visual cues. Example approaches include the analysis of replays [10, 12, 44], crowd cheering [6, 36], motion features [35], and closed captioning [40]. More recently, Bettadapura et al. [7] used contextual cues from the environment to understand the excitement levels within a basketball game. Tang and Boring [32] proposed to automatically produce highlights by analyzing social media services such as twitter. Decroos et al. [8] developed a method for forecasting sports highlights to achieve more effective coverage of multiple games happening at the same time. Different from existing methods, our proposed approach offers a unique combination of excitement measures extracted from live video streams to produce highlights, including information from the *spectators*, the *commentator*, and the *player* reaction. As such, our system incorporates and combines most of the information employed by previous works (audio, visual, text). It could also be easily extended to integrate other sources of attention or excitement, such as social media feeds or production cues (replays, closed captions, etc.). In addition, we enable personalized highlight generation or retrieval based on a viewer's favorite players.

Self-Supervised Learning. In recent years, there has been significant interest in methods that learn deep neural network classifiers without requiring a large amount of manually annotated training examples. In particular, *self-supervised* learning approaches rely on auxiliary tasks for feature learning, leveraging sources of supervision that are usually available "for free" and in large quantities to regularize deep neural network models. Examples of auxiliary tasks include the prediction of ego-motion [4, 13], location and weather [34], spatial context or patch layout [22, 24], image colorization [43], and temporal coherency [21]. Aytar et al. [5] explored the natural synchronization between vision and sound to learn an acoustic representation from *unlabeled* video. We leverage this work to build audio models for crowd cheering and commentator excitement using few training examples, and use those classifiers to constrain the training data collection for player reaction recognition. More interestingly, we exploit the detection of TV graphics as a free supervisory signal to learn feature representations for player recognition from unlabeled video.

III. Technical Approach

A. Framework

Our framework is illustrated in Figure 2. Given an input video feed, we extract in parallel multimodal markers of potential interest: player action of celebration and facial expression (detected by visual classifiers), crowd cheer (with an audio classifier), commentator excitement (detected by a combination of an audio classifier and a salient keywords extractor applied after a speech-to-text component), and game analyst input information when available (text based metadata). The start and end of a potential highlight clip are determined via analyst input when it is available. In the absence of such input, the start of a highlight is determined by identifying graphic content overlaid to the video feed signifying the start of a shot. Similarly, the end of a highlight segment is identified with visual shot boundary detection, applied in a window of few seconds after the occurrence of the last excitement marker. Additionally, by applying an OCR engine to the graphic, we can recognize the name of the player involved as well as additional metadata such as the hole number, nature of the shot, etc. Finally we compute a combined excitement score for the segment proposal based on a combination of the individual markers. In the following we describe each component in detail.

B. Audio-based Markers Detection

1) Crowd Cheer Detection: Crowd cheering is perhaps the most veritable form of approval of a player's shot within the context of any sport. Cheers almost always accompany important shots. Most critically, crowd cheer can point to the fact that an great point or shot was just played (indicating the end of a highlight). Another important audio marker is excitement in the commentators' tone while describing a shot. Together those two audio markers play a key role in determining the position and excitement level of a potential highlight clip. We leverage SoundNet [5] to construct audio-based classifiers for crowd-cheering and commentator tone excitement. Soundnet uses a deep 1-D convolutional neural network architecture to learn representations of environmental sounds from nearly 2 million unlabeled videos. The classes learned by SoundNet are objects and scenes, and they do not include crowd cheering or clapping, not excitement tone in a persons voice. Therefore a domain adaption step needs to be performed in order to use such powerful representation for our purposes. Instead of fine-tuning two SoundNet models, one for the specific task of crowd cheering classification and one for commentator excitement tone classification, we chose to employ the same SoundNet deep features as basis to train a linear SVM model for each of the two markers. We opted for this approach for two reasons. The first is the relatively limited amount of training data available for both tasks. We wanted to limit the amount of annotation effort needed to build efficient and effective models. In practice, when dealing with medium or small scale training data, the literature is not conclusive on whether fine-tuning a deep network is better than learning another model (such as SVM) on top of deep features [26]. The second is efficiency. While one could argue that a multi-task fine-tuned network could have achieved the same result, we picked a simpler solution. We extract features from the *conv-5* layer in SoundNet to represent 6 seconds audio segments. The choice of the *conv-5* layer is based upon experiments and superior results reported in [5]. The dimensionality of the feature is 17,152. We then learn a linear SVM model atop the deep features to classify crowd cheer.

We adopt an iterative refinement bootstrapping methodology to construct our audio based classifiers. We learn an initial classifier with relatively few audio snippets (28 positives and 57 negatives in the first round of training) and then perform a few rounds of bootstrapping on a distinct set. This procedure is repeated to improve the accuracy at each iteration. Cheer samples from 2016 Masters replay videos as well as examples of cheer obtained from YouTube were used as positive training data. For negative examples, we used audio tracks containing regular speech, music, and other kinds of non-cheer sounds found in Masters replays. In total our final training set consisted of 156 positive and 193 negative samples (6 seconds each). The leave-one-out cross validation accuracy on the training set was 99.4%.

2) Commentator Excitement Detection: We propose a novel commentator excitement measure based on a combination of voice tone and speech-to-text-analysis.

Tone-based: Besides recognizing crowd cheer, we employ the deep Soundnet audio features to model excitement in commentators' tone. As above, we employ a linear SVM classifier for modeling. For negative examples, we use audio tracks containing regular speech, music, regular cheer (without commentator excitement) and other kinds of sounds which do not have an excited commentator. In total, the training set for audio based commentator excitement recognition consisted of 131 positive and 217 negative samples. The leave-one-out cross validation accuracy on the training set was 81.3%.

Text-based: While tone can say a lot about how excited the commentator is while describing a shot, excitement level can also be gauged from another source, that is, the expressions used. We created a dictionary of 60 expressions (words and phrases) indicative of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2876046, IEEE Transactions on Multimedia
The final version of record is available at http://dx.doi.org/10.1109/TMM.2018.2876046

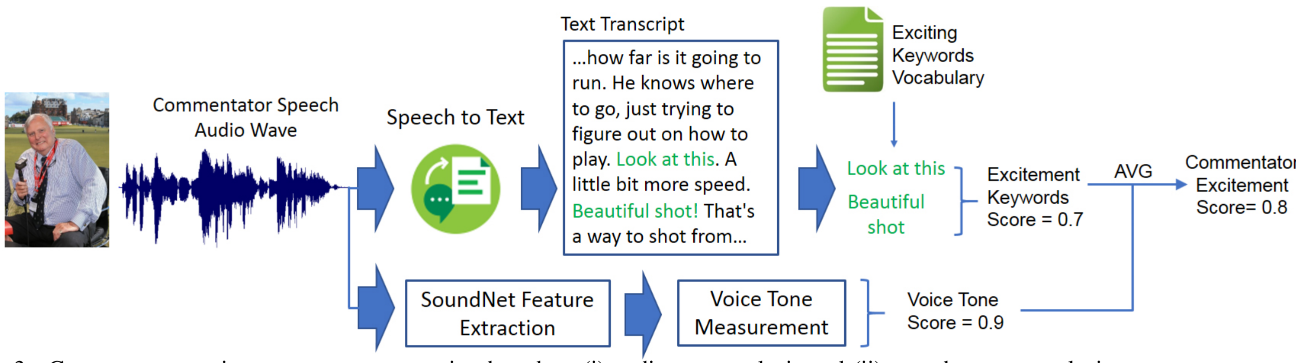


Fig. 3. Commentator excitement score computation based on (i) audio tone analysis and (ii) speech to text analysis.

excitement (e.g. "great shot", "fantastic") and assign to each of them excitement scores ranging from 0 and 1. We use a speech to text service[1] to obtain a transcript of commentators' speech and create an excitement score as an aggregate of scores of individual expressions in it.

Finally we average the tone-based and text-based scores to obtain the overall level of excitement of the commentator, as exemplified in Figure 3.

C. Visual Marker Detection

1) Player Reaction: Understanding the reaction of a player is another important cue to determine an interesting moment of a game. In our work, we train an action recognizer to detect a player celebrating. To the best of our knowledge, measuring excitement from the player reaction for highlight extraction has not been explored in previous work. We adopt two strategies to reduce the cost of training data collection and annotation for action recognition. First, we use our audio-based classifiers (crowd cheer and commentator excitement) at a low threshold to select a subset of video segments for annotation, as in most cases the player celebration is accompanied by crowd cheer and/or commentator excitement. Second, inspired by [17], we use still images which are much easier to annotate and allow training with less computational resources compared to video-based classifiers. Figure 4 shows examples of images used to train our model. 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. Since ImageNet does not contain categories describing a person celebrating, a fine-tuning procedure for our specific domain is needed. We collect Positive training examples for the fine-tuning from 2016 Masters, Wimbledon, and US Open videos, and also from the web. Negative examples are randomly sampled from the aforementioned videos. Similarly to the audio models, multiple rounds of bootstrapping were used to train the model. Details of the training procedure are described in the Section V-B.

Fig. 4. Examples of still images used to train our player celebratory action recognition model.

2) Facial Expression: 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, thus confusing the player reaction model. 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 [16]. We retrieved "face" bounding boxes when detected, or "head-shoulders" ones alternatively. When multiple boxes were detected, we selected only the largest one. We also ignored poor examples due to occlusion, rear-angle, or partial visibility. The extracted faces were then categorized into four types of expression: aggressive, tense, smiling, and neutral, as shown in Figure 5. The first three are associated with celebration, whereas the last one is considered as non-celebratory. The facial expression classifier was trained by fine-tuning a VGG-face [23] model on a manually labeled dataset of tennis players faces.

[1] *https://www.ibm.com/watson/developercloud/speech-to-text.html*

3) TV Graphics, OCR, and Shot-boundaries: In professional golf tournament broadcasts, a golf swing is generally preceded by a TV graphics with the name of the player just about to hit the golf ball and other information about the shot. The detection of such markers is straightforward, as they appear in specific locations of the image, and have distinct colors. We check for such colors in the vicinity of pre-defined image locations (which are fixed across all broadcasting channels) to determine the TV graphics bounding box. One could use a more general approach by training a TV graphics detector (for example via Faster-RCNN [27] or SSD [16]), however this was beyond the scope of this work. We apply OCR (using the Tesseract engine [30]) within the detected region to extract metadata such as the name of the player and the hole number. This information is associated with the detected highlights, allowing personalized queries and highlight generation based on a viewers favorite players. We then use standard shot-boundary detection based on color histograms as a visual marker to determine the end of a highlight clip.

D. Game Analytics

In tennis not every point, for how exciting it may be, 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 statisticians positioned on the side of the court, 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:

- -0.1 for a point won due to unforced error or rally count smaller than 3
- +0.1 for a point won due to positive play, volley winner, smash winner, match point, break point, or rally count greater than 5
- +0.20 for a point won due to forced error, player movement detected, or rally count greater than 10
- +0.25 for a game winning point

where positive play means a point won thanks to a player's active effort, not an opponent's mistake. Player movement signifies that one player moved 25 meters more than the opponent. The sum of values for any given point was then normalized in the range 0 to 1.

E. Excitement Scores Fusion

For for any given potential highlight clip x, we perform late fusion of the excitement scores $E_n(x)$ produced by each of the N marker classifiers. Specifically, we aggregate (via the max operation) positive scores for each of the markers within the inspected time-window (usually of 15-20 seconds). Each individual score is then registered in the range between 0 and 1 via sigmoid normalization, and the final fusion is computed as a weighted linear sum:

$$F(x) = \sum_{n=1}^{N} w_n E_n(x) \quad (1)$$

where n refers to $cheer$, $commentator$, player $action$ and game $analytics$ (when available). The weights w_n for each component are learned via cross-validation on data from the previous year's tournaments. Crowd cheer, commentator excitement (combining audio and text), player reaction and game analytics components weights were set to 0.61, 0.26, 0.13 and 0 respectively for Masters. For Wimbledon and US Open they were learned as 0.6, 0, 0.1, and 0.3 respectively. In Section V-C1 we compare the benefit of learning the weights versus a Naive-Fusion approach employing equal weighting across components.

Fig. 5. Examples of different expressions used to train the facial expression model.

F. Highlight Detection

A highlight is identified as a play or shot that receives a high overall score from the fusion score combining the multimodal markers ones. Besides measuring marker response, it is also important to determine the start and end positions of a highlight. This step is handled differently for the two use cases of golf and tennis, since the inputs to the system were different. We will go through them individually.

Golf: in this case, the input to the system are the live video streams of the 2017 Masters. Figure 6 illustrates then how we incorporate multimodal markers to identify segments as potential highlights and assign excitement scores to them. The system starts by generating *segment proposals* based on the crowd cheering marker. Specifically, crowd cheering detection is performed on a continuous segment of the stream and positive scores are tapped to point to potentially important cheers in audio. Adjacent 6 second segments with positive scores are merged to mark the end of a bout of contiguous crowd cheer. Each distinct cheer marker is then evaluated

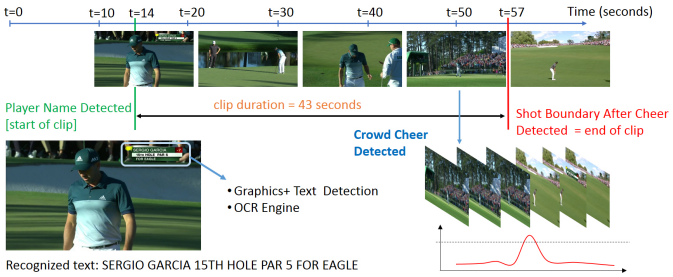


Fig. 6. Highlight clip start and end frames selection pipeline for the golf video streams.

as a potential candidate for a highlight using presence of a TV graphics marker containing a player name and hole number within a preset duration threshold (set at 80 seconds). The beginning of the highlight is set as 5 seconds before the appearance of TV graphics marker. In order to determine the end of the clip we perform shot boundary detection in a 5 second video segment starting from the end of cheer marker. If a shot boundary is detected, the end of the segment is set at the shot change point. Segments thus obtained constitute valid highlight segment proposals for the system.

Tennis: As opposed to the golf Masters, Wimbledon and US Open tennis matches are actively annotated by analysts in a live fashion. As a consequence, the start and end times of each play can be accurately determined based on such provided information. Therefore the multimodal marker classification system receives video clips filtered using analyst information as potential highlight candidates and ranks them.

TV graphics detection, shot boundary detection and OCR could be applied to tennis in the same way as they were applied to golf. As our system is motivated by application to real world production needs, we did not investigate the clip detection and cut approach to the tennis videos, since the clips and the player information were already provided to us during the tennis tournaments. However, we believe it would apply seamlessly, as the TV Graphics are easily identifiable and OCR could be employed to identify player names and keep track of the points. The only needed addition would be that of a player serving detection marker, since the start of a tennis point clip corresponds to one player serving. That would require training a specific classifier, which could be done similarly to the player celebratory reaction one, without requiring a big effort.

For both golf and tennis use cases, highlight clips are displayed in the system dashboard as shown in Figure 1 labeled with individual marker scores (normalized between 0 and 1) as well as a combined excitement score which is computed as a linear combination of the multimodal marker scores.

IV. Self-Supervised Player Recognition

Automatic player detection and recognition can be a very powerful tool for generating personalized highlights when graphics are not available, as well as to perform analysis outside of the event broadcast itself. It could for example enable to estimate the presence of a player in social media posts by recognizing his face. The task is however quite challenging. First, there is a large variations in pose, illumination, resolution, occlusion (hats, sunglasses) and facial expressions, even for the same player, as visible in Figure 11. Second, inter-player differences are limited, as many players wear extremely similar outfits, in particular hats in golf, which occlude or obscure part of their face. Finally, a robust face recognition model requires large quantities of labeled data in order to achieve high levels of accuracy, which is often difficult to obtain and labor intensive to annotate. We propose to alleviate such limitations by exploiting the information provided by other modalities of the video content, specifically the overlaid graphics containing the players name. This allows us to generate a large set of training examples for each player, which can be used to train a face recognition classifier, or learn powerful face descriptors. In the following we describe the approach employed specifically for the golf tournament data, but it could be easily adapted to other sports.

We start by detecting faces within a temporal window after a graphic with a player name is found, using a Faster-RCNN detector [27]. The assumption is that in the segment after the name of a player is displayed, his face will be visible multiple times in the video feed. Not all detected faces in that time window are going to represent the player of interest. We therefore perform outliers removal, using geometrical and clustering constraints. We assume the distribution of all detected faces to be bi-modal, with the largest cluster containing faces of the player of interest. Faces that are too small are discarded, and faces in a central position of the frame are given preference. Each face region is expanded by 40% and rescaled to 224x224 pixels. Furthermore, only a maximum of one face per frame can belong to a given player. Given all the face candidates for a given player, we perform two-class k-means clustering on top of fc7 features extracted from a VGG Face network [23], and keep only the faces belonging to the largest cluster while respecting the geometric constraints to be the representative examples of the player's face. This process, working without supervision, allows us to collect a large quantity of training images for each player. We can then train a player face recognition model, which in our case consists of a VGG Face Network fine-

tuned by adding a softmax layer with one dimension per player. Figure 11(b) shows an example subset of training faces automatically collected for Sergio Garcia from the 2016 Masters broadcast. The system was able to collect hundreds of images with a large variety of pose and expressions for the same player. Bordered in red are highlighted two noisy examples. While the purity of the training clusters is not perfect, as we will show in the experiments of Section V-D it still allowed to learn a robust classifier with no explicit supervision. This confirms recent results in deep learning modeling, which has been proven to being robust to noise if a large quantity of training data is provided, as demonstrated in recent results for example by Veit et al. on OpenImages [33] or recently achieved top ImageNet performance using noisy data from image tags by Mahajan et al. [19].

V. Experiments

A. Experimental Setting

We evaluated our system in three real world championships, namely the 2017 Masters, 2017 Wimbledon, and 2017 US Open tournaments. For the 2017 Masters, we analyzed in near real-time the content of four channels broadcasting simultaneously over the course of four consecutive days, from April 6th to April 9th, for a total of 124 hours of content[2]. Our system produced 741 highlights over all channels and days. The system ran on a Redhat Linux box with two K40 GPUs. We extracted frames directly from the video stream at a rate of 1fps and audio in 6 seconds segments encoded as 16bit PCM at rate 22,050 kHz. The cheer detector and commentator excitement ran in real time (1 second to process one second of content), action detection took 0.05secs per frame, graphics detection with OCR took 0.02secs per frame. Speech-to-text was the only component slower than real time, processing 6 seconds of content in 8 seconds, since we had to upload every audio chunk to an API service. The 2017 Wimbledon and US Open system ran on two Ubuntu nodes with four K80 GPUs each, providing a total of 16 stream services to process candidate highliht clips during the tournaments. Videos were chunked in 10 seconds clips and analyzed in less than 2.5 seconds through our service APIs. Frames and audio extracted from each video were distributed to several components for crowd cheering detection, action recognition, expression recognition, and overall aggregation.

In the following we report experiments conducted after the events to quantitatively evaluate the performance of H5, both in terms of overall quality of the produced highlights as well as efficacy of its individual components. All training was performed on content from the 2016 Masters, Wimbledon and US Open tournament videos and from images downloaded from the web, while testing was done on video data from the 2017 tournaments.

[2]Video replays are publicly available at *http://www.masters.com/en_US/watch/index.html*

B. Individual Markers

We first present evaluation of individual markers as performed on golf and tennis data.

1) Player Celebration Marker: The player celebration classifier for 2017 Masters was trained with 574 positive examples and 563 negative examples. The positive examples were sampled from 2016 Masters replay videos and also from the web. The negative examples were randomly sampled from the 2016 Masters videos. We used the VGG-16 model [28], pre-trained on Imagenet as our base model. The Caffe [14] deep learning library was used to fine-tune the model to our data with stochastic gradient descent, learning rate 0.001, momentum 0.9, and weight decay 0.0005. We performed three rounds of hard negative mining on 2016 Masters videos, obtaining 2,906 positive examples and 6,744 negative ones.

In a similar fashion, player celebration classifiers for 2017 Wimbledon and US Open were trained using samples from 2016 tournament video frames as well as examples from the web including multiple rounds of bootstrapping. The final training set for Wimbledon consisted of 13,263 positive and 33,372 negatives samples, augmented by random cropping and horizontal flipping. Since the US open setting is quite different from Wimbledon's, we trained new celebration classifiers for US Open using a training set consisting of 11,330 positive and 12,516 negative samples. The examples from the web were reused from the Wimbledon training, while new video frames were annotated specifically for the US Open. We explored two deep architectures, VGG-16 and ResNet-50 pre-trained on Masters data, and found the ResNet model to work best.

We evaluated the player celebration models on a set of clips randomly selected from each of the tournaments and manually labeled. Table I reports the details of each test set. 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. Compared to VGG-16, the ResNet-50 models performed better on 2017 Wimbledon (AUC = 0.91 versus 0.87 for VGG) while they were equivalent for the US Open (both AUC = 0.94). Because of the superior

Event	# clips	# frames	# positives	# negatives
2017 Masters	-	1,064	59	1,005
2017 Wimbledon	540	4,777	78	4,699
2017 US Open	510	8,963	52	8,911

TABLE I
DETAILS OF THE TEST SETS USED TO EVALUATE THE PLAYER CELEBRATION ACTION RECOGNITION MODELS.

Event	# samples	# positives	# negatives
2017 Masters	405	69	336
2017 Wimbledon	1,073	915	158
2017 US Open	1,564	627	937

TABLE III
DETAILS OF THE TEST SET USED TO EVALUATE PERFORMANCE OF THE CROWD CHEERING RECOGNITION MODELS.

Event	# samples	#aggressive	# tense	# smiling	# neutral
2017 Masters	1,285	45	222	346	672
2017 Wimbledon	472	18	56	38	360
2017 US Open	1,129	25	171	44	889

TABLE II
DETAILS OF THE TEST SET USED TO EVALUATE PERFORMANCE OF THE FACIAL EXPRESSION RECOGNITION MODELS.

Event	# samples	# positives	# negatives
2017 Masters	240	46	194
2017 Wimbledon	437	85	352
2017 US Open	423	111	312

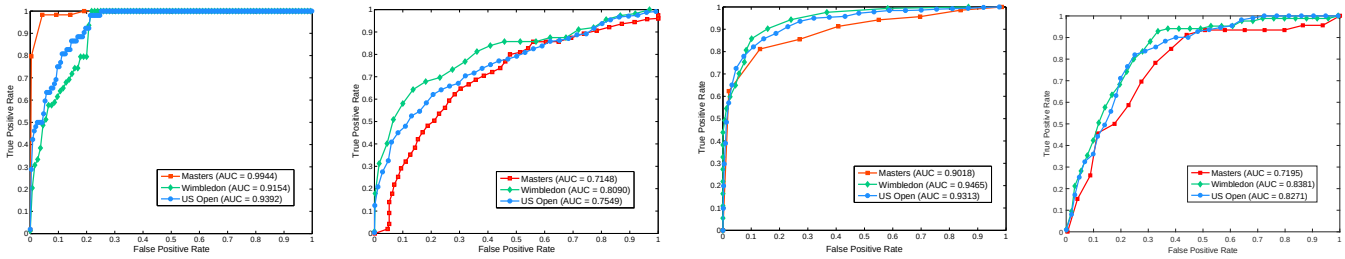
TABLE IV
DETAILS OF THE TEST SET USED TO EVALUATE PERFORMANCE OF THE COMMENTATOR TONE RECOGNITION MODELS.

performance of ResNet, we used those models in the fusion phase. In Figure 7(a) we show the ROC curves of the best models for the all three inspected tournaments. We can observe that recognizing players celebrating was easier in golf than in tennis. In fact, despite a significantly smaller training dataset, the model performs better. This is mostly due to false positives occurring in tennis when players serve, catch a ball in their hand, or pass a towel over their head. The high false positives rate from the the player reaction models was one of the main motivations to introduce a facial expression module.

2) Facial Expression Marker: The facial expression marker was tested on faces extracted from the 2017 Wimbledon and US Open test set videos. While it was not employed during the tournament, we also evaluated this marker on the 2017 Masters data after the event. As shown in the Table II, the expressions on the players faces were at first manually labeled into four categories, which we found to be most representative of the appearance of the players from the 2016 tournaments. During the 2017 Wimbledon however, we found that aggressive, tense and smiling all correlated with players' celebrations. We therefore combined those facial expressions with a linear fusion to generate an overall "excited" score, which was compared against the neutral score representing a lack of celebration. Figure 7(b) illustrates the ROC curves of the facial expression marker using this binary categorization. The Figure shows that the modules performed reliably enough for both tennis tournaments, with 2017 Wimbledon's performance being better (AUC of 0.81 for 2017 Wimbledon versus 0.75 for 2017 US Open). Recognition accuracies are 82.42% and 82.23%, respectively. The results for golf were worse, with AUC of 0.71 and accuracy of 78.57%. While the performance is not by itself perfect, it resulted in being acceptable since facial expressions were used to refine the results given by the celebration action model.

3) 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, Wimbledon and US Open tournaments videos to evaluate the performance of the model. Table III reports detailed numbers of test sets and Figure 7(c) reports performance of the audio cheer model. The resulting ROC curves are approximately similar, with AUC of 0.9, 0.94 and 0.93 for 2017 Masters, Wimbledon and US Open, respectively.

4) Commentator Tone Marker: In a fashion similar to identifying crowd cheering samples, we created a training set for commentator tone excitement marker using 2016 Masters videos and several rounds of bootstrapping. In total, the training set for audio based commentator excitement recognition consisted of 131 positive and 217 negative samples. The model was employed only for the golf Masters tournament, as the tennis data we were provided as input did not contain any useful commentary. We evaluated the commentator tone model on randomly sampled audio snippets from the 2017 Masters tournament, and from Youtube videos of past Wimbledon and US Open matches, as summarized in Table IV. Each audio clip was sent to AMT for evaluation by 5 different workers, who had to label it as *No speech* (0), *Softly spoken*(1), *Average Excitement*(2), *Loud Excitement*(3). Any clip with an average score of at least 2 was considered as exciting, while the others were



(a) Player Reaction (b) Facial Expression (c) Crowd Cheer (d) Commentator Tone

Fig. 7. ROC curves for different excitement markers on the test clips from the 2017 Masters, Wimbledon and US Open tournaments.

Element	Total Number	Precision	Recall
Words	7,663	0.9916	0.9893
Characters	29,016	0.9846	0.9840

TABLE V
OCR PERFORMANCE IN TERMS OF WORDS AND CHARACTERS RECOGNITION.

considered non-exciting. Figure 7(d) shows the ROC curves of the model, with an AUC = 0.72, 0.83 and 0.82 for Masters, Wimbledon and US Open, respectively. 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 Master tournament. It was not used for the tennis ones.

5) Text OCR Marker: In order to evaluate the text detection and OCR performance, we randomly selected 625 frames from the 4 channels during the first day of the 2017 Masters tournament. For each of the frames, we manually transcribed the ground truth text and compared it to the outcome of the OCR engine. From the results in Table V we observe that the system was able to recognize the overlaid text very accurately. Overall, only in 7 frames the name of the player was not properly recognized, while the most common mistake (happened 60 times) was the confusion of the letter *T* with the letter *I* in the ordinal numbers indicating the hole (for example, *15TH HOLE* misspelled as *15YH HOLE*). Precision and Recall are computed as

$$Precision = \frac{N_{cor}}{N_{gt}}, \quad Recall_c = \frac{N_{cor}}{N_r} \tag{2}$$

$$N_{cor} = N_{gt} - ED\,(s_g, s_r) \tag{3}$$

where N_{cor} is the number of correctly recognized characters, that is, the number N_{gt} of ground truth characters minus the edit distance $ED\,(s_g, s_r)$ between the ground truth text s_g and the text output from the system s_r.

C. Highlights Detection

Evaluating the quality of sports highlights is a challenging task, since a clearly defined ground truth does not exist. Similarly to previous works [7], we approached this problem by comparing the clips automatically generated by our system to two human based references. The first is a human evaluation and ranking of the clips that we produced. The second is the collection of highlights professionally produced by the official Masters curators and published on their Twitter channel.

1) Human Evaluation of Highlights Ranking: In order to determine the quality of the rankings produced by our system, we conducted user studies on Amazon Mechanical Turk. Workers were asked to evaluate the excitement level of several clips randomly sampled from the ones generated and scored by the H5 framework. We asked each participant to assign a score to every clip in a scale from 0 to 5, with 0 meaning a clip without any interesting content and 5 being the most exciting shots. We then averaged the scores of the users for each clip. Table VI summarizes the number of clips and workers employed for each tournament. We also asked each worker if they were fans of the given sport, and on average we found half of them being fans.

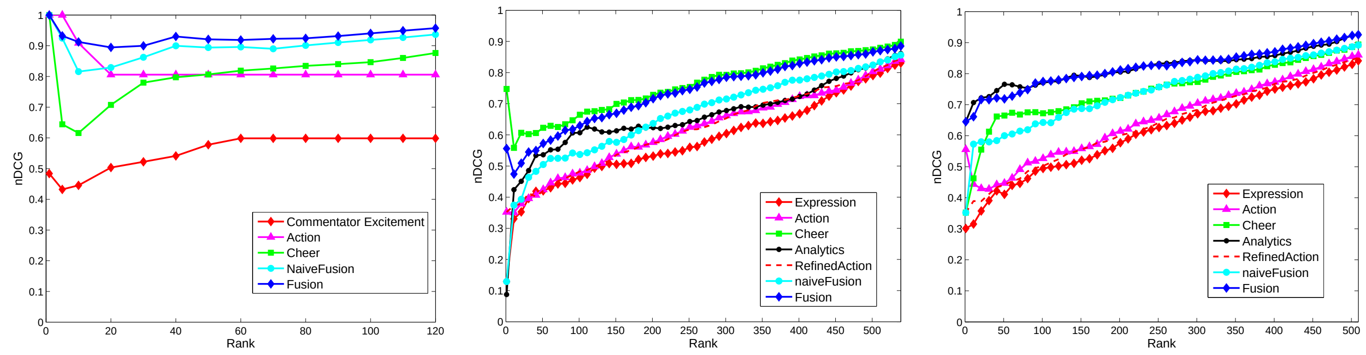
Specifically for 2017 Masters, a score of 1 had the unique meaning of a highlight that is associated with the wrong player, that is, a system mistake. The resulting scores determined that 92.68% of the clips produced by our system were legitimate highlights (scores 2 and above), while 7.32% were mistakes.

We then compared the rankings of the clips according to the scores of each individual component, as well as their fusion, to the ranking obtained through the users votes. The performance of each ranking was computed at different depth k with the normalized discounted cumulative gain (nDCG) metric, which is a standard retrieval measure computed as follows

$$nDCG(k) = \frac{1}{Z} \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2(i+1)} \tag{4}$$

where rel_i is the relevance score assigned by the users to clip i and Z is a normalization factor ensuring that the perfect ranking produces a nDCG score of 1.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2876046, IEEE Transactions on Multimedia
The final version of record is available at http://dx.doi.org/10.1109/TMM.2018.2876046



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

Fig. 8. nDCG computed at different ranks for the individual components as well as the Fusion.

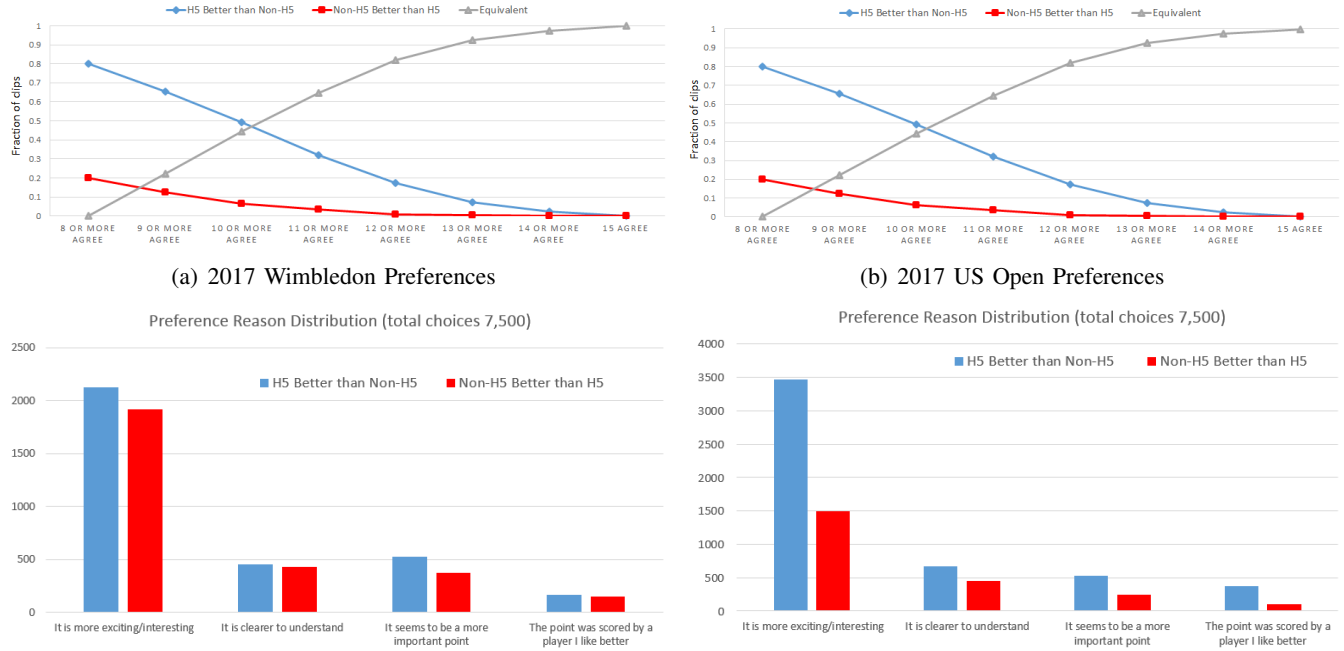
In Figure 8 we present the nDCG at different ranks. Fusion was obtained as a weighted sum of the normalized scores from each component. We tested two different fusion configurations: a Naive-Fusion using equal weights, and the Fusion with weights optimized through cross-validation on a separate training set (which was used by the system during the tournaments), as described in Section III-E. In all cases the system's Fusion outperforms the Naive one, confirming the benefit of assigning different weights to different individual components. For both 2017 Wimbledon and US Open, the Refined Action obtained by adjusting the player celebration score with the facial expression component (increase if facial expression is "excited", decrease if the expression is neutral) correlated better than the base Action one, confirming the benefit of introducing the facial expression marker. In general, H5 produced rankings which correlated more with human preferences for Golf than for Tennis. For 2017 Masters (a) we notice that all components but the Commentator Excitement correctly identify the most exciting clip (at rank 1). After that only the Action component assigns the highest scores to the following top 5 clips. When considering 10 top clips or more, the benefit of combining multiple modalities becomes apparent, as the Fusion nDCG curve remains constantly higher than each individual marker. Differently from 2017 Masters, for 2017 Wimbledon and US Open the Fusion does not outperform individual components. However it remains fundamental for the system to generalize, as it is interesting to notice how for different tournaments, different components correlated most with human rankings. For Wimbledon (b) Cheer was the best indicator for human excitement, whereas for US Open (c) the game Analytics mattered the most. In both cases the Fusion closely follows the performance of the best individual marker.

2) Tennis A/B Testing: Besides the ranking of clips, for Tennis we also wanted to determine whether the

Tournament	Number of Clips	Workers per Clip	Total N. Workers	% Fan Workers
2017 Masters	120	3	3	33%
2017 Wimbledon	540	5	33	58%
2017 US Open	510	5	21	59%

TABLE VI
DISTRIBUTION OF CLIPS AND WORKERS USED TO EVALUATE CLIPS RANKINGS FOR EACH TOURNAMENT.

selection made by the system about which clips should go into the compiled highlights and which should be instead discarded followed human preferences. Thus we evaluated the clip selection process through another Amazon Mechanical Turk experiment. In this case 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 then presented each pair to the workers and asked them to pick which clip in the pair was more exciting and/or interesting. We also asked the workers to motivate their choice among multiple options and to provide some demographic information. Each pair was voted on by 15 workers, and a total of 234 unique users participated in the study. From the results reported in Figure 9 (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. The distribution of reasons for the choices is highly skewed toward how exiting a clip was, as users paid less attentions to clip clarity (tends to be very similar, as clips belong to the same game), players scoring or significance of a point



(a) 2017 Wimbledon Preferences

(b) 2017 US Open Preferences

(c) 2017 Wimbledon Reasons

(d) 2017 US Open Reasons

Fig. 9. Human preferences in AB tests for 2017 Wimbledon and US Open.

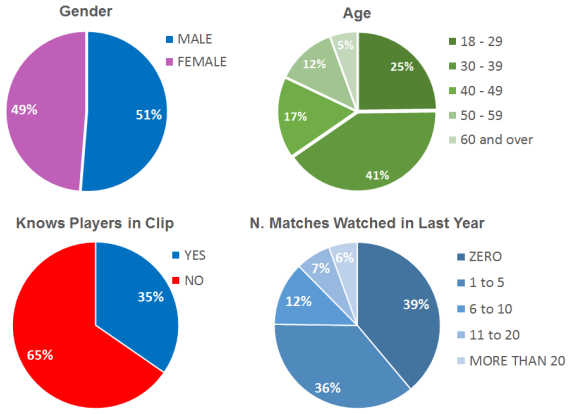


Fig. 10. A/B Tests users demographics information

withing a game. The detailed breakdown is presented in Figure 9 (c) and (d). From the demographic information collected in Figure 10 we can observe a quite even distribution in gender, with a prevalence of young people (18 to 29 years old) who mostly did not know the players in the clips they voted on. This is consistent with the reason *the point was scored by the player I like better* being the least used in Figure 9 (c) and (d). Finally, it seems that the majority of workers was not a tennis fan, having watched less than 5 games in the past year.

3) Comparison with Official 2017 Masters Highlights: The previous experiments confirmed the quality of the identified highlights as perceived by potential users of the system. We then compared H5 generated clips with highlights professionally created for 2017 Masters, *Masters Moments*, available at their official Twitter page[3]. There are a total of 116 highlight videos from the final day at the 2017 Masters. Each one covers a player's approach to a certain hole (e.g. Daniel Berger, 13th hole) and usually contains multiple shots taken to complete a particular hole. In contrast each H5 highlight video is about a specific shot at a particular hole for a given player. In order to match the two sets of videos, we considered just the player names and hole numbers and ignored the shot numbers. After eliminating Masters Moments outside of the four channels we covered live during the tournament and for which there is no matching player graphics marker, we obtained 90 Masters Moments.

In Table VII, we report Precision and Recall of matching clips over the top 120 highlights produced by the H5 Fusion system. We observe that approximately half of the clips overlap with Masters Moments. This leaves us with three sets of videos: one shared among the two sets (a gold standard of sorts), one unique to Masters Moments and one unique to H5. We observed that by lowering thresholds on our markers detectors, we can incorporate 90% of the Masters Moments by producing more clips. Our system is therefore potentially capable of producing almost all of the professionally produced content. We also wanted to investigate the quality of the clips which were discovered by the H5 system beyond what the official Master's channel produced. Generation of highlights is a subjective task and may not

[3] *https://twitter.com/mastersmoments*

Depth	120	500
Precision	0.54	0.35
Recall	0.4	0.9
Matching Highlights Preference	0.57	-
Non-Matching Highlights Preference	0.33	-
Equivalent	0.10	-

TABLE VII
HIGHLIGHTS DETECTION PERFORMANCE. COMPARISON BETWEEN THE TOP k ($k = 120, 500$) RETRIEVED CLIPS FROM H5 AND THE OFFICIAL 2017 MASTER'S TWITTER HIGHLIGHTS.

comprehensively cover every player and every shot at the Masters. At the same time, some of the shots included in the official highlights may not necessarily be great ones but strategically important in some ways.

While our previous experiment was aimed at understanding the coverage of our system vis-a-vis official 2017 Masters highlights, we wondered if a golf aficionado would find the remaining videos still interesting (though not part of official highlights). We therefore aimed an experiment at quantitatively comparing (a) H5 highlight clips that matched Masters Moments and (b) H5 highlight clips that did not match Masters Moments videos. In order to do so we selected the 40 most highly ranked (by H5) videos from lists (a) and (b) respectively and performed a user study using three human participants familiar with golf. Participants were shown pairs of videos with roughly equivalent H5 scores/ranks (one from list (a) and the other from list (b) above) and were asked to label the more interesting video between the two, or report that they were equivalent. Majority voting was used among the users votes to determine the video pick from each pair. From the results reported in Table VII we observe that while the preference of the users lies slightly more for videos in set (a), in almost half of the cases the highlights uniquely and originally produced by the H5 system were deemed equally if not more interesting. This reflects that the system was able to discover content that users find interesting and goes beyond what was officially produced. It is also interesting to notice that our system is agnostic with respect to the actual score action of a given play, that is, a highlight is detected even when the ball does not end up in the hole, but the shot is recognized as valuable by the crowd and/or commentator and players through their reactions to it.

D. Self-Supervised Player Face Recognition

In order to test our self-supervised player recognition model we randomly selected a set of 10 players who participated to both the 2016 and the 2017 Masters tournaments (shown in Figure 11 (a)). In Table VIII we

Fig. 11. Self-supervised player face learning. (a) Examples of the 10 players used in the experiments. (b) Subset of the images automatically selected as training set (2016 Masters) for Sergio Garcia (note the diversity of pose, expression, occlusion, illumination, resolution). (c) Examples of test faces (2017 Masters) correctly recognized through self-supervised learning. (d) Examples of False Negatives (in orange) and False Positives (in red).

report the statistics of the number of training images that the system was able to automatically obtain in a self-supervised manner. For each player we obtain on average 280 images. Data augmentation in the form of random cropping and scaling was performed to uniform the distribution of examples across players. Since there is no supervision in the training data collection process, some noise in bound to arise. We manually inspected the purity of each training cluster (where one cluster is the set of images representing one player) and found it to be 94.26% on average. Note that despite evaluating its presence, we did not correct for the training noise, since our method is fully self-supervised. The face recognition model was fine-tuned from a VGG-face network with learning rate = 0.001, $\gamma = 0.1$, momentum = 0.9 and weight decay = 0.0005. The net converged after approximately 4K iterations with batch size 32. We evaluated the performance of the model on a set of images randomly sampled from Day 4 of the 2017 Msters and manually annotated with the identity of the 10 investigated players. Applying the classifier directly to the images achieved 66.47% accuracy (note that random guess is 10% in this case since we have 10 classes). We further clustered temporally close frames based on fc7 features and assigned to all faces in a cluster the identity which received the highest number of predictions within the cluster. This process raised the performance to 81.12%. Figure 11 (c) shows examples of correctly labeled test images of Sergio Garcia. Note the large variety of pose, illumination, occlusion and facial expressions. In row (d) we also show some examples of false negatives (bordered in orange) and false positives (in red). The net result of our framework is thus a

Number of Players	10
Number of Training Images	2,806
Training Clusters Purity	94.26%
Number of Test Images	1,181
Random Guess	10.00%
Classifier Alone Accuracy	66.47%
Classifier + Clustering Accuracy	**81.12%**

TABLE VIII
2017 MASTERS PLAYER FACE CLASSIFICATION PERFORMANCE.

self-supervised data-collection procedure which allows to gather large quantities of training data without need for any annotation, which can be used to learn robust feature representations and face recognition models.

E. Discussion

Ablation study results. The combination of multi-modal excitement measures is crucial to determine the most exciting moments of a game. Though crowd cheer is an important marker, it alone cannot differentiate a hole-in-one or the final shot of a golf tournament from other equally loud events. In addition, we noticed several edge cases where non-exciting video segments had loud cheering from other holes. Our system correctly attenuates the highlight scores in such cases, due to the lack of player celebration and commentator excitement. In tennis, we observed how the player celebration marker can produce false positives associated with raising one's hands for purposes other than celebrating (for example cleaning one's sweat from the forehead). Our system copes with it by analyzing the player's facial expression in conjunction with his or her actions.

Comparison to the state of the art and extensions. Many state of the art approaches for sports and video analytics are actually complementary to ours. We believe that other sources of excitement measures, such as as replays [10, 12], crowd facial expressions or information from social media feeds [32] could be easily integrated within our framework to further improve it. The live feed nature of the video streams we analyzed during the tournaments, which are the input of our system, made it impossible to rely on production cues such as replays for our purposes at the time. Similarly we did not have access to social media feeds during the events. Integrating such complimentary cues could be a good direction for future work. Also, end-to-end approaches to video description or action recognition (to further capture the plasticity of a move, for example) could be employed within our framework, although currently the lack of large-scale annotated training data hinders the development of such approaches. Finally, most existing works utilize one or a subset of the components we employ within our framework. For example Baijal et al. [6] and Xiong et al. [35] use audio events (such as crowd cheering) only, Zhang et a. [41] employ closed captions analysis, which can be equated to the commentator text analysis we perform on the output of the speech to text module. As such, the extensive ablation study we performed with the evaluation of the contribution of each individual component in our framework and their combination, as reported in Section V-C, can be considered as a proxy for comparison with many existing state of the art methods.

Other uses of self-supervised learning. The same approach used for self-supervised player recognition could also be applied for the detection of other items, for example golf setup (player ready to hit the golf ball), tennis player serving or handshake at the end of a game, using TV graphics or other modalities metadata as a proxy to obtain positive examples without manual supervision. This would generalize our approach to detect the start of an event without relying on TV graphics, and also help fix a few failure cases of consecutive shots for which a single TV graphics is present.

Extension to other sports. While we have demonstrated our approach for golf and tennis, we believe our proposed techniques for modeling the excitement levels of the players, commentator, and spectators are general and can be extended to other sports as well since most of our markers are sport agnostic. However, it should be noted that both tennis and golf are relatively quiet sports, where exciting events are rare. A sport like basketball or soccer has the crowds chanting all the time and it would be challenging to directly employ a completely sport-agnostic system like ours out-of-the-box, without any adaptation. In those instances, specialized knowledge of the sport in question can definitely add value to the highlight selection process. An integration of sport-specific action detection markers (i.e. a basketball dunk, or a soccer goal) might be helpful for the system to work. On the other hand, the system might be directly applicable to similar quiet sports such as cricket.

VI. CONCLUSION

We presented a novel approach for automatically extracting highlights from sports videos based on multi-modal sport-independent excitement measures, including audio analysis from the spectators and the commentator, and visual analysis of the players. Based on that, we developed a first-of-a-kind system for auto-curation of golf and tennis highlight packages, which was demonstrated in three major golf and tennis tournaments in 2017. We also exploited the correlation of different modalities to learn models with reduced cost in training

data annotation. As next steps, we plan to generalize our approach to other sports such as soccer and produce more complex storytelling video summaries of the games, while including additional indicators of play importance such as social media feeds or game specific events detection.

REFERENCES

[1] http://www.zdnet.com/article/ibm-watson-is-creating-highlight-reels-at-the-masters/.
[2] https://www.cnet.com/news/ibm-wimbledon-highlights-artificial-intelligence/.
[3] https://www.nytimes.com/2017/09/05/sports/us-open-highlights.html/.
[4] P. Agrawal, J. Carreira, and J. Malik, "Learning to see by moving," in *ICCV*, 2015.
[5] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *NIPS*, 2016.
[6] A. Baijal, J. Cho, W. Lee, and B.-S. Ko, "Sports highlights generation based on acoustic events detection: A rugby case study," in *ICCE*, 2015.
[7] V. Bettadapura, C. Pantofaru, and I. Essa, "Leveraging contextual cues for generating basketball highlights," in *ACM Multimedia*, 2016.
[8] T. Decroos, V. Dzyuba, J. Van Haaren, and J. Davis, "Predicting soccer highlights from spatio-temporal match event streams," in *AAAI*, 2017.
[9] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.
[10] T. Hasana, H. Boril, A. Sangwan, and J. H. L. Hansen, "Multi-modal highlight generation for sports videos using an information-theoretic excitability measure," *EURASIP Journal on Advances in Signal Processing*, 2013.
[11] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Automatic trailer generation," in *ACM Multimedia*, 2010.
[12] A. Javed, K. B. Bajwa, H. Malik, and A. Irtaza, "An efficient framework for automatic highlights generation from sports videos," *Signal Processing Letters*, vol. 23, no. 7, 2016.
[13] D. Jayaraman and K. Grauman, "Learning image representations tied to ego-motion," in *ICCV*, 2015.
[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimedia*, 2014.
[15] 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.
[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *ECCV*, 2016.
[17] 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.
[18] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *ACM Multimedia*, 2002.
[19] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten, "Exploring the limits of weakly supervised pretraining," in *ECCV*, 2018.
[20] 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.
[21] H. Mobahi, R. Collobert, and J. Weston, "Deep learning from temporal coherence in video," in *ICML*, 2009.
[22] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*, 2016.
[23] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.
[24] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016.
[25] A. Rav-Acha, Y. Pritch, and S. Peleg, "Making a long video short: Dynamic video synopsis," in *CVPR*, 2006.
[26] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *CVPRW*, 2014.
[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.
[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[29] J. R. Smith, D. Joshi, B. Huet, W. Hsu, and J. Cota, "Harnessing a.i. for augmenting creativity: Application to movie trailer creation," in *ACM Multimedia*, 2017.
[30] R. Smith, "An overview of the tesseract ocr engine," in *ICDAR*, 2007.
[31] M. Sun, A. Farhadi, and S. Seitz, "Ranking domain-specific highlights by analyzing edited videos," in *ECCV*, 2014.
[32] A. Tang and S. Boring, "# epicplay: Crowd-sourcing sports video highlights," in *ACM CHI*, 2012.
[33] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie, "Learning from noisy large-scale datasets with minimal supervision," in *CVPR*, 2017.
[34] J. Wang, Y. Cheng, and R. Feris, "Walk and learn: Facial attribute representation learning from egocentric video and contextual data," in *CVPR*, 2016.
[35] Z. Xiong, R. Radhakrishnan, and A. Divakaran, "Generation of sports highlights using motion activity in

combination with a common audio feature extraction framework," in *ICIP*, 2003.

[36] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework," in *ICASSP*, 2003.

[37] H. Xu, Y. Zhen, and H. Zha, "Trailer generation via a point process-based visual attractiveness model," in *IJCAI*, 2015.

[38] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, and B. Guo, "Unsupervised extraction of video highlights via robust recurrent auto-encoders," in *ICCV*, 2015.

[39] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in *CVPR*, 2016.

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

[41] J. Zhang, J. Yu, and D. Tao, "Local deep-feature alignment for unsupervised dimension reduction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2420–2432, 2018.

[42] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *ECCV*, 2016.

[43] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *ECCV*, 2016.

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