IBM High-Five: Highlights From Intelligent Video Engine

Dhiraj Joshi[1], Michele Merler[1], Quoc-Bao Nguyen[1], Stephen Hammer[2], John Kent[2], John R. Smith[1], Rogerio S. Feris[1]
[1]IBM T. J. Watson Research Center,[2] IBM iX
{djoshi,mimerler,quocbao,hammers,johnkent,jsmith,rsferis}@us.ibm.com

ABSTRACT

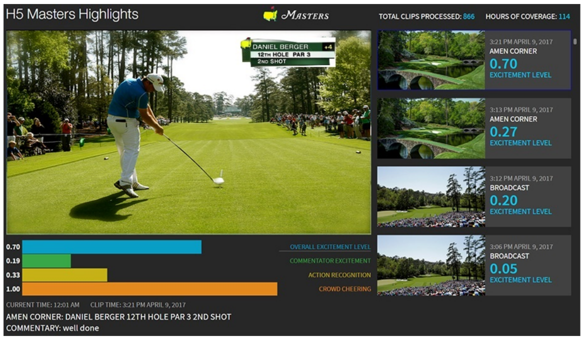
We introduce a novel multi-modal system for auto-curating golf highlights that fuses information from players' reactions (celebration actions), spectators (crowd cheering), and commentator (tone of the voice and word analysis) to determine the most interesting moments of a game. The start of a highlight is determined with additional metadata (player's name and the hole number), allowing personalized content summarization and retrieval. Our system was demonstrated at Masters 2017, a major golf tournament, generating real-time highlights from four live video streams over four days.

CCS CONCEPTS

• **Information systems → Multimedia content creation**; *Personalization*;

KEYWORDS

Sport Analytics; Multimodal Video Analysis; Highlights Generation



TOTAL CLIPS PROCESSED: 866 HOURS OF COVERAGE: 114

3:21 PM APRIL 9, 2017
AMEN CORNER
0.70
EXCITEMENT LEVEL

3:13 PM APRIL 9, 2017
AMEN CORNER
0.27
EXCITEMENT LEVEL

3:12 PM APRIL 9, 2017
BROADCAST
0.20
EXCITEMENT LEVEL

3:06 PM APRIL 9, 2017
BROADCAST
0.05
EXCITEMENT LEVEL

0.70 OVERALL EXCITEMENT LEVEL
0.19 COMMENTATOR EXCITEMENT
0.33 ACTION RECOGNITION
1.00 CROWD CHEERING

CURRENT TIME: 12:01 AM CLIP TIME: 3:21 PM APRIL 9, 2017
AMEN CORNER: DANIEL BERGER 12TH HOLE PAR 3 2ND SHOT
COMMENTARY: well done

Figure 1: The IBM H5 system dashboard for auto-curation of sports highlights generated in near real-time (right panel). Users can click on the icons on the right panel to play the associated video in the center, along with the scores for each excitement measure.

1 INTRODUCTION

Generation of sports highlights is a manual and labor-intensive effort that often requires condensing content from hundreds of hours of video into a few minutes of key defining moments. Moreover many viewers prefer personalized highlights such as the collection of best shots from their favorite player(s).

In this demonstration, we will present a novel system for auto-curating golf highlights that combines multi-modal information from multiple sources, i.e. the *player*, *spectators*, and the *commentators* to determine a game's most exciting moments in near real-time. The highlights are added to an interactive dashboard where they can be potentially reviewed by a video editor, thus speeding up the highlight generation and sharing process. Figure 1 shows the interface of our system, called High-Five (**High**lights from **I**ntelligent **V**ideo **E**ngine), H5 in short. In addition, by automatically extracting metadata via TV graphics and OCR, we allow personalized highlight retrieval or alerts based on player name, hole number, location, and time.

2 FRAMEWORK

Our framework is illustrated in Figure 2. Given an input video feed, we extract in parallel four multimodal markers of potential interest: player action of celebration (detected by a visual classifier), 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).

2.1 Multimodal Markers

Crowd-cheer: In this work, we leverage SoundNet[1] to construct an audio-based classifier for crowd-cheering. Soundnet uses a deep 1-D convolutional neural network architecture to learn representations of environmental sounds from nearly 2 million unlabeled videos. We learn a linear SVM model atop the deep features to classify crowd cheer.

Commentator-excitement: We propose a novel commentator excitement measure based on voice tone and speech-to-text-analysis. **Tone-based:** We employ the deep SoundNet audio features (as for crowd cheer) to model excitement in commentators' tone. Similar to crowd cheer, we employ a linear SVM classifier for modeling commentator tone excitement. **Text-based:** We create a dictionary of 60 expressions (words and phrases) indicative of excitement (e.g. "great shot", "fantastic") and assign to each of them excitement scores. We then use the IBM Speech to Text Service[1] to obtain speech transcripts and aggregate scores of individual expressions in it.

Player-celebration: We train an image based model to recognize a player celebrating. We use the VGG-16 model[3] pre-trained on Imagenet as our base model, and further fine tune it on a set of player celebration images obtained by sampling and annotating frames temporally close to segments identified by our audio-based crowd cheer classifier.

MM '17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-4906-2/17/10. .
https://doi.org/10.1145/3123266.3127924

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

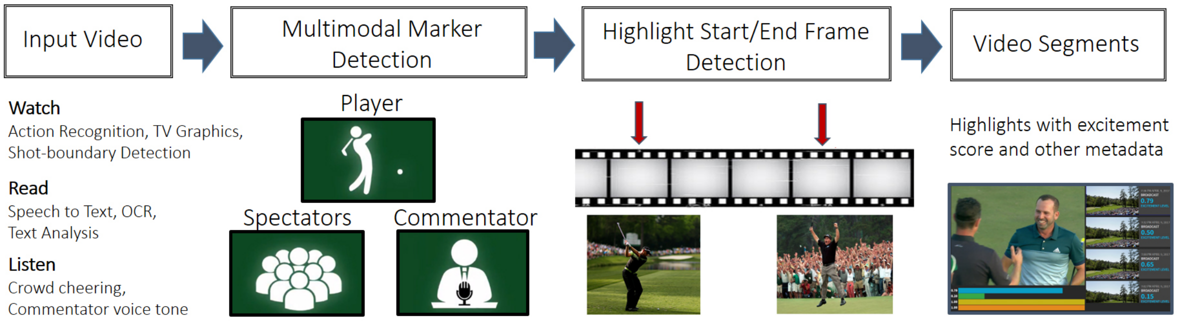

Input Video
Multimodal Marker Detection
Highlight Start/End Frame Detection
Video Segments
Watch
Action Recognition, TV Graphics, Shot-boundary Detection
Read
Speech to Text, OCR, Text Analysis
Listen
Crowd cheering, Commentator voice tone
Player
Spectators
Commentator
Highlights with excitement score and other metadata

Figure 2: Framework of our system. Multimodal (video, audio, text) marker detectors measure the excitement levels of the player, spectators, and commentators to generate video segment highlights.

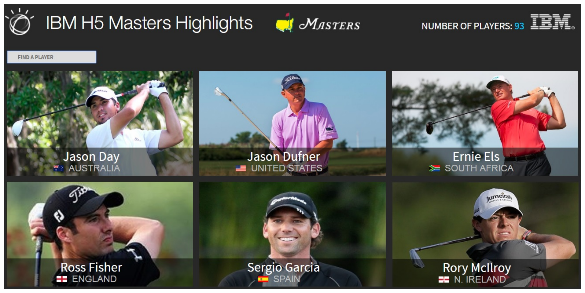


Figure 3: Personalized player based highlights selection dashboard.

2.2 Highlight Detection

The system starts by generating **segment proposals** based on the crowd cheering marker identifying potential moments of excitement. 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 segments with positive scores are merged to mark the end of a bout of contiguous crowd cheer. The start of the segment is identified by player graphics content overlaid to the video feed. We apply an OCR engine to the graphics to recognize the name of the player involved and the hole number, as well as additional metadata about the shot. The end of the segment is determined using a visual shot boundary detection applied in a window of a few seconds after the occurrence of the last cheer marker in the segment. Finally we compute a combined excitement score for the segment proposal based on fusion from multi-modal markers.

3 SYSTEM DESCRIPTION

Live Demonstration at Masters 2017: Our system was demonstrated live in a real world major Golf tournament (Golf Masters 2017). The system analyzed in near real-time the content of the 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. Our system produced 741 highlights over all channels and days. The system back end performed the core processing in near real-time and posted highlight segment information in *json* format. This was received by the front end which generated and posted videos on the highlight dashboard (Figure 1).

System Back End: The back end system runs on a Redhat Linux box with two K40 GPUs. Frames are extracted directly from the video stream at a rate of 1fps and audio in 6 seconds segments encoded as 16bit PCM at rate 22,050. The audio is given to the crowd cheer and commentator excitement classifiers that run in real time (1 second to process 1 second of content). Frames are given to the player celebration action detector that takes 0.05secs per frame. Graphics detection with OCR takes 0.02secs per frame. The speech-to-text is the only component slower than real time, processing 6 seconds of content in 8 seconds, since we have to upload every audio chunk to an API service. Individual components process their input in parallel and finally the fusion component creates the segment proposals and fuses scores from multiple markers to generate the final highlights. All the models were trained on content from the 2016 Golf Masters broadcast. Further technical details as well as system evaluation are presented in [2].

System Front End: Highlights are displayed with an associated excitement level score at the system front end (as shown in Figure 1). Users can click on the icons on the right panel to play the associated highlight in the center while excitement scores corresponding to individual markers are also displayed. The system also allows users to retrieve personalized highlights by players (as shown in Figure 3) and rank highlights by individual modalities.

4 DEMONSTRATION

At the conference, we will demonstrate the IBM H5 dashboard used live at Masters 2017. We will invite all interested attendees to interact with our dashboard, view the system-created highlights, and experience the system in person. We look forward to interaction and valuable user-feedback by multimedia experts at the conference.

REFERENCES

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. SoundNet: Learning Sound Representations from Unlabeled Video. In *NIPS*.

[2] Michele Merler, Dhiraj Joshi, Quoc-Bao Nguyen, Stephen Hammer, John Kent, John R. Smith, and Rogerio S. Feris. 2017. Automatic Curation of Golf Highlights using Multimodal Excitement Features. In *CVPR Int. Workshop on Sports*.

[3] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).