What You Hear Is What You Get: Audio-Based Video Content Analysis

Benjamin Elizalde, Gerald Friedland
International Computer Science Institute
1947 Center Street
Berkeley, CA 94704, USA
benmael, fractor@icsi.berkeley.edu

Karl Ni
Lawrence Livermore National Laboratory
Livermore, CA 94550
ni4@llnl.gov

Abstract

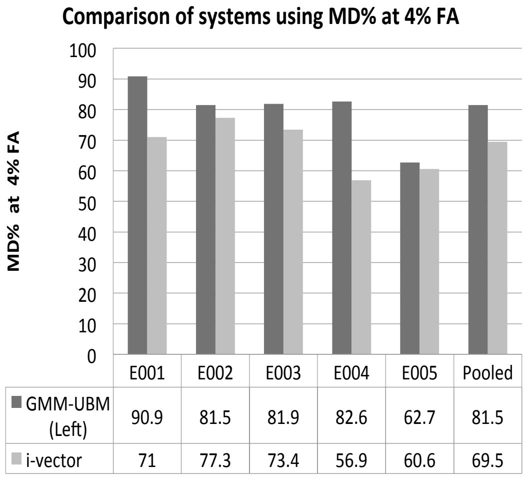
Audio-based video event detection on user-generated content (UGC) aims to find videos that show an observable event, such as a wedding ceremony or a birthday party. In a lower tier, audio concept detection aims to find a sound or concept, such as music, clapping or a cat's meow. Different events are described by different sounds. The difficulty of video content analysis on UGC lies in the lack of structure and acoustic variability of the data. The video content analysis has been explored mainly by computer vision, but it requires audio to complement the search of cues on this multimedia challenge. This paper presents an approach for each detection task. First, an i-vector system for audio-based video event detection. The system compensates for undesired acoustic variability and extracts information from the acoustic environment of the event recordings, making it a meaningful choice for event detection on UGC. Second, an audio concept ranking-based neural network system that aids to determine and select the most relevant concepts for each event, to discard meaningless concepts, and to combine ambiguous sounds to enhance a concept.

1 Video Content Analysis

A brief description of our two different systems for the video content analysis is included, along with the main challenges and latest results.

1.1 Video Event Detection

Video event detection on user-generated content (UGC) aims to identify videos with a semantically defined event, such as a wedding ceremony or birthday party rather than an object, such as a wedding dress, or an audio concept, such as clapping or laughter. The difficulty of video content analysis on UGC lies in the lack of structure and the variability of contents from videos in the same event categories. The video event detection system employed is an i-vector based system [1]. The system involves creating a low-dimensional vector characterizing the acoustic event of each video's audio track, referred to as the i-vector. The strength of the algorithm is that it takes into account the within- and between-event acoustic scatter, allowing the algorithm to account for scenarios where multiple videos of the same event have different acoustic characteristics, and where videos from different events have similar acoustic characteristics. Therefore, the technique provides a valid approach not only for tackling the event detection task itself, but also for handling the difficulties of UGC data. The research includes a performance comparison with a conventional GMM-UBM-based. The i-vector system outperforms the GMM-UBM-based system, in terms of the Missed Detection (MD) rate at 4% by an absolute 12% as shown in Figure 1



	E001	E002	E003	E004	E005	Pooled
■ GMM-UBM (Left)	90.9	81.5	81.9	82.6	62.7	81.5
■ i-vector	71	77.3	73.4	56.9	60.6	69.5

Figure 1: The MD at 4% FA performances of the GMM-UBM, and i-vector systems, for five individual event categories and the pooled category (E001-E005: Attempting a Board Trick, Feeding an Animal, Landing a Fish, Wedding Ceremony, Working on a Woodworking Project).

Table 1: The ranking-based selection achieved the best overall mean accuracy, and the highest Term Frequency-Inverse Document Frequency's event relevance for the top 40 audio concepts.

Selection based on	Mean accuracy	TF-IDF score
Highest-accuracy	20.38 %	0.83
Frame-frequency	18.33 %	0.90
Ranking	21.55 %	1

1.2 Audio Concept Detection

Audio concepts such as music, speech, clapping or water running define an acoustic fingerprint that defines a video and distinguishes it from its cohorts. Different event categories are better described by different concepts; therefore, audio concept classification provides evidence of the video's event. The difficulties of this task are subjective annotations, concept ambiguities, overlap, short duration, and prominence among others. The audio concept detection system [2] is based on a Neural Network approach because it has demonstrated high performance on a similar task where it discriminates well between different sounds called phonemes. The audio extracted from the video feeds the system, which outputs for each audio frame, predictions for each of the labeled concepts. Afterwards a confidence threshold on the predictions is defined, hence providing reliable evidence of concepts for each video. The research includes a performance and relevance to events comparison 1 of using a selection of audio concepts based on the number/amount of frames in the annotations, another selection based on the highest-accuracy concepts and lastly a selection based on the proposed ranking methodology, which aids to select the best trade-off between relevance to the event and frame classification accuracy.

References

[1] Benjamin Elizalde, Howard Lei, Gerald Friedland, Capturing the acoustic scene characteristics for Audio Scene Detection, *in submission to IEEE International Symposium on Multimedia ISM 2013*.

[2] Benjamin Elizalde, Mirco Ravanelli, Gerald Friedland, Audio Concept Ranking for Video Event Detection on User-Generated Content, *in Proceedings of Interspeech's Workshop on Speech, Language and Audio in Multimedia SLAM 2013*.