Audio-Concept Features and Hidden Markov Models for Multimedia Event Detection

Benjamin Elizalde[1]*, Mirco Ravanelli*[2]*, Karl Ni*[3]*, Damian Borth*[1]*, Gerald Friedland*[1]

[1]International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704, USA
[2]Fondazione Bruno Kessler, via Sommarive 18, 38123 Trento, Italy
[3]Lawrence Livermore National Laboratory, Livermore, USA
{benmael,borth, fractor}@icsi.berkeley.edu, mravanelli@fbk.eu, karl_ni@llnl.gov

Abstract

Multimedia event detection (MED) on user-generated content is the task of finding an event, e.g., a *Flash mob* or *Attempting a bike trick*, using its content characteristics. Recent research has focused on approaches that use semantically defined "concepts" trained with annotated audio clips. Using audio concepts allows us to show semantic evidence of their relationship to events, by looking at the probability distribution of the audio concepts per event. However, while the concept-based approach has been useful in image detection, audio concepts have generally not surpassed the performance of low-level audio features like Mel Frequency Cepstral Coefficients (MFCCs) in addressing the unstructured acoustic composition of video events. Such audio-concept based systems could benefit from temporal information, due to one of the intrinsic characteristics of audio: it occurs across a time interval. This paper presents a multimedia event detection system that uses audio concepts; it exploits the temporal correlation of audio characteristics for each particular event at two levels. The first level involves analyzing the short- and long-term surrounding context information for the audio concepts, through an implementation of a Hierarchical Deep Neural Network (H-DNN), to determine engineered audio-concept features. At the second level, we use Hidden Markov Models (HMMs) to describe the continuous and non-stationary characteristics of the audio signal throughout the video. Experiments using the TRECVID MED 2013 corpus show that an HMM system based on audio-concept features can perform competitively when compared with an MFCC-based system.

Index Terms: Audio Concepts, Video Event Detection, TRECVID MED, Deep Neural Networks, Hidden Markov Models.

1. Introduction

Web videos, also called user-generated content (UGC), are the fastest-growing type of content on the Internet. To make this data accessible, we must be able to automatically organize and analyze the content of recordings. In this paper, we investigate experimental methods for multimedia event detection (MED), the process of finding videos that relate to a semantically defined event, such as *Flash mob*, *Dog show*, or *Town hall meeting*. This task is implicitly multimodal, in that events in video are characterized by audio-visual cues. Depending on the user query, the audio stream can sometimes be more descriptive than the visual stream. For instance, audio cues can allow one to quickly determine certain characteristics of the environment, the tone of the video, or the presence of music.

There have been several approaches to MED; nevertheless, the research has increasingly revolved around semantic or humanly explainable approaches. As a result, the focus has shifted toward concept detection [1, 2, 3, 4], where the visual domain has shown relative success. However, analysis in the audio domain still relies mainly on low-level features [5, 6], because they significantly outperform audio concepts for event detection. Reasons for this gap in performance are varied. For example, the audio concepts defined must actually discriminate between events to provide reliable detection. Additionally, UGC often presents mismatched conditions, which limits the performance of the trained concept detectors with unseen audio. Nonetheless, based on the idea that video events could be described by sets of different audio concepts such as *water running* or *metallic clanking*, audio concept detection continues to be explored, for example in [7, 8, 9]. However, approaches sometimes ignore the intrinsic temporal characteristics of audio, even though it has been suggested that it helps, for example in [10], where local discriminant bases (low-level features) are used to feed an HMM and exploit temporal correlation of environments. Moreover, in [11], [12], researchers created an HMM-GMM model for every audio concept, to get a segmentation based on the transitions between the models, rather than a fixed segmentation. All in all, MED field requires systems that use audio concepts to show event evidence and at the same time provide competitive performance with low-level features.

This paper presents a multimedia event detection system that uses audio concepts. It is a follow-up to our research presented in [13]. The use of audio concepts allows us to show semantic evidence of their relationship with events, by looking at the probability distribution of the audio concepts per event. However, this system exploits the temporal correlation of audio characteristics for a particular event at two levels. The first level involves analyzing the short and long-term surrounding context information for audio concepts, through an implementation of a Hierarchical Deep Neural Network (H-DNN), to determine engineered audio-concept features. At the second level, we use Hidden Markov Models (HMMs) to describe the continuous and non-stationary characteristics of the audio signal throughout the video. The presented method is evaluated on the TRECVID MED 2013 corpus, showing that an HMM-system based on audio-concept features can be competitive in comparison to an MFCC-based system.

The paper is structured as follows. Section 2 describes the

video dataset and the audio concepts we used in the experiments. Section 3 details the MED system using audio concepts. Section 4 analyzes the experimental results. Lastly, Section 5 summarizes our conclusions and suggests future work.

2. The TRECVID MED Corpus

The video dataset used in the experiments is from the NIST TRECVID Multimedia Event Detection 2013 corpus, composed of 150,000 UGC videos with an average length of three minutes [14]. We used the EK100 set, consisting of 100 videos each for 20 event categories (listed in Table 2) and the MED Test set. The audio in those videos is unstructured and contains environmental acoustics, overlapping sounds, and unintelligible audio, among other characteristics.

The audio concepts we used are listed in Table 1; these 40 concepts were selected in previous research by some of the authors as having the highest event-relevance and detection performance [13]. The selection was made from a pool of audio concepts corresponding to three different sets of annotations, by Carnegie Mellon University (CMU) [15], SRI & Stanford University (STAND) [8], and SRI-Sarnoff (SRI) [4]. The length of the audio concepts varies from a fraction of a second to a few seconds. The labels defined start time and end time. Some of the audio-concept labels are similar (*anim[al] bird* (CMU), *birds* (STAND)), or at least closely semantically related (*speech* (CMU), *conversational speech* (SRI)). However, it is important to note that the annotation projects had different procedures and used different sets of event-videos, so similarity of labels does not necessarily mean the labeled concepts have identical acoustic characteristics; we therefore did not merge them. (In Section 4.1, we show that this choice is justified by the experimental data.)

#	Concept	Source	#	Concept	Source
1	engine light	CMU	21	crowd yells	STAND
2	wind	STAND	22	speech not eng.	CMU
3	speech	CMU	23	crowd laughter	STAND
4	metallic clanking noises	SRI	24	word 'tire' spoken	SRI
5	crowd cheers	STAND	25	rolling	SRI
6	washboard	CMU	26	anim bird	CMU
7	water running	STAND	27	singing	SRI
8	children's voices	SRI	28	crowd	CMU
9	airtraffic	STAND	29	mumble	CMU
10	birds	STAND	30	music	CMU
11	small party	STAND	31	environmental	STAND
12	water splashing	STAND	32	laughing	SRI
13	audio of wedding vows	SRI	33	water	CMU
14	rustle	CMU	34	processed	CMU
15	crowd applause	STAND	35	individual yells	STAND
16	conversational speech	SRI	36	singing	CMU
17	radio	CMU	37	crowd noise	SRI
18	scratch	CMU	38	squeak	CMU
19	micro blow	CMU	39	noise of passing cars	SRI
20	engine quiet	CMU	40	board hitting surface	SRI

Table 1: The set of 40 audio concepts annotated by CMU, STAND, and SRI with the highest event-relevance and detection performance.

3. The Multimedia Event Detection System Using Audio Concepts

This section describes the audio concept–based MED system shown in Fig. 1, and how we implemented it to test performance.

3.1. The Audio Concept Detection System

The audio concept detection system, which computes the audio-concept features to feed the HMM-based event detector, is based on the H-DNN system proposed in [16], which performs well at detecting audio concepts in UGC.The audio from the event videos is extracted and MFCCs of 13 coefficients are computed using a 25 ms Hamming window with a stride size of 10ms per frame shift. After a mean-and-variance normalization step, a context window is applied to gather 49 consecutive frames. We chose MFCCs as the main feature type because of their established performance in audio tasks. Prior to the input layer of the first neural network (NN), a dimensionality reduction step using a Discrete Cosine Transform (DCT) is applied to de-correlate the information provided by the context window. The DCT step selects relevant information, and has shown benefits in contrast with a non-DCT step [16]. Lastly, the features are fed to the H-DNN.

The H-DNN architecture is a cascade of two neural networks. As discussed in [16], [17], this configuration is particularly effective because it analyses both the short and long-term modulations of each audio concept (performed by the first and the second NN, respectively), which significantly improves detection accuracy. Most audio concepts, including music, clapping, knocking, laughing, and many others, are characterized by several iterations of a similar pattern over the time span, thus making short and long-term analysis of the audio concepts effective. In this case, the first-level NN in Fig. 1(a), which converts the low-level features into a higher-level representation, is composed of 3 hidden layers with 2000 neurons per layer. Because it has more than 2 hidden layers, pre-training based on the Restricted Boltzann Machine (RBM) [18] is adopted for initialization and convergence. The output layer, a softmax-based classifier, outputs a 40-dimensional vector, corresponding to the number of audio concepts.

To achieve a long-term analysis of the audio-concept modulations, we sample the features generated by the NN depicted in Fig. 1(a) at the 5 positions: -10, -5, 0, +5, +10. The long-term NN depicted in Fig. 1(b) is composed of 2 hidden layers with 1000 neurons each. Compared to that in Fig. 1(a), the NN in Fig. 1(b) thus employs a shallower architecture; we chose this configuration because the task of the long-term NN is less convoluted. That is, the MLP in Fig. 1(a) realizes a conversion from low-level to high-level features, while Fig. 1(b) operates on already-processed input streams from the previously trained NN.

The NN training and pre-training phases used TNet [19]. Pre-training initializes weights in the first two hidden layers via RBM (Gaussian-Bernoulli) using a learning rate of 0.005 with 10 pre-training epochs. The remaining RBMs (Bernoulli-Bernoulli) use a learning rate of 0.05 with 5 pre-training epochs.

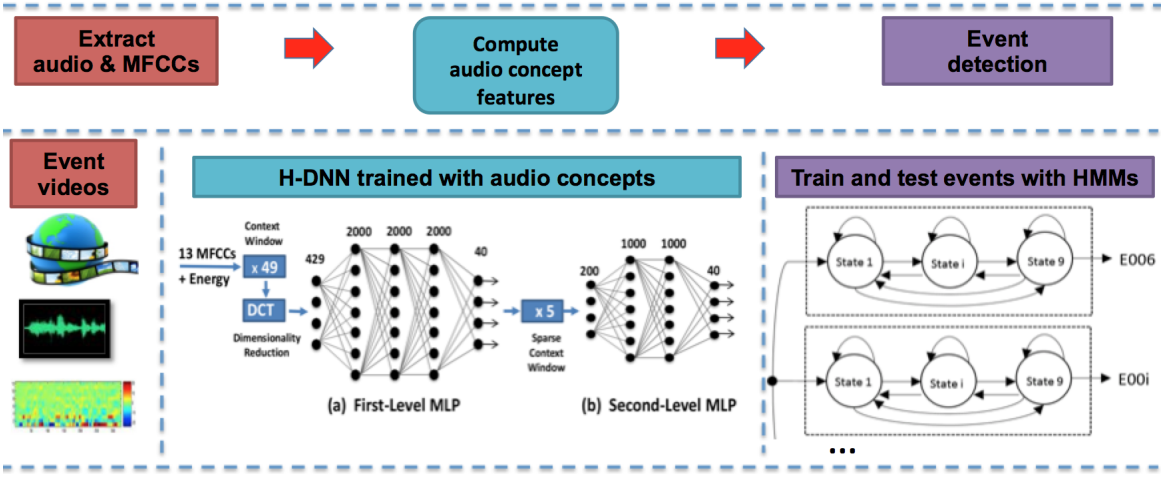

Figure 1: The MED system using audio concepts extracts the audio track from the event videos. It then computes MFCCs from the audio as the input to the Hierarchical Deep Neural Network. The H-DNN ouputs the audio-concept features, which are fed to the event-trained HMM-based event detection system. Lastly, an event-detection decision is made for the given test video.

From the training set, we derived a small cross-validation set (10% of the training data) for the following back-propagation training. The fine-tuning phase is performed by a stochastic gradient descent that optimizes the cross-entropy loss function. The learning rate, updated by the "new-bob" algorithm, is kept fixed at 0.002 as long as the single-epoch increment in the cross-validation frame accuracy is higher than 0.5%. For subsequent epochs, the learning rate is halved until the cross-validation increment in accuracy is less than the stopping threshold of 0.1%. NN weights and biases are updated per blocks of 1024 frames.

3.2. The Event Detection System

We propose an event detection approach based on Hidden Markov Models (HMM). An HMM is a generative statistical model; the technique is widely used in temporal pattern-detection problems such as gesture recognition, genomics, and handwriting, as well as speech recognition [20]. In an HMM, an observable variable x, which represents the input features, is assumed to be generated by a sequence of internal hidden states S, which cannot be observed directly. An HMM is completely described by the number of hidden states, the transition probabilities between hidden states, the initial state probabilities, and the probability distribution of each possible state $P(x/S)$. The later probability density function is usually modeled by exploiting a Gaussian Mixture Model (GMM) composed of a certain number of gaussians. The HMMs' parameters are derived in the training phase by means of the Baum-Welch algorithm, while the pattern detection is usually carried out using the Viterbi algorithm, as seen in [20].

One of the most interesting features of an HMM is its inherent capacity to detect complex temporal patterns, which in this paper is exploited to detect events based on their audio tracks. The adopted HMM topology is depicted in Fig. 1. In speech recognition, usually only forward connections between states are used to model the well-defined temporal evolution of each phone; in contrast, we propose using fully-connected models, because videos of events rarely follow any well-defined time evolution (for instance, in a *Dog show* video we can have music, laughing, clapping, and barking in different sequences). The training and test phases were carried out using the HTK [21].

4. Results and Analysis

In this section, we describe the performance of the proposed multimedia event detection system using audio concepts, and discuss how the distribution of the audio concepts can support a description of the events.

4.1. Audio Concepts Distribution

We analyze the distribution of audio concepts using the output of the H-DNN system—in other words, the audio-concept features. For each event, we took the semantic features of the corresponding 100 training videos and add all the posterior probabilities for each of the 40 audio concepts, to obtain an audio-concept probability histogram for each of the 20 video events. The most common audio concepts detected across all events were the ones related to speech, music, and crowds. The least-often detected audio concepts have the characteristics of being quite or short, such as *engine light (soft)*, *micro blow* (a puff of breath, such as one might use to blow out a single candle), and *squeak*.

Example histograms showing the audio concepts corresponding to three examples are shown in Fig. 2. For *Birthday party*, the audio concepts with the highest probabilities are *small party STAND* (a small group of people celebrating), *conversational speech SRI*, *singing SRI*, and *processed CMU*. (The concept *processed CMU* refers to music added to the video in post-processing.) Spot-checking the videos, we found that these audio concepts match people speaking and singing the "birthday song", and sometimes music in the background (though not necessarily added in post-processing). In another example, *Winning a race without a vehicle* corresponds to the audio concepts *small party STAND*, *crowd cheers STAND*, *crowd Noise SRI*, *crowd CMU*, and *wind STAND*. Spot-checking the videos, we find that the most common sounds are indeed crowd-related. Lastly, *Grooming an animal* corresponds to *speech CMU*, *small party STAND*, *water splashing STAND*, *conversational speech SRI*, *water running STAND*, and *processed CMU*. Spot-checking the videos, we find people describing how to groom the animal, a variety of water sounds (splashing, running, coming out of a hose, etc.), and, less often, user-added music.

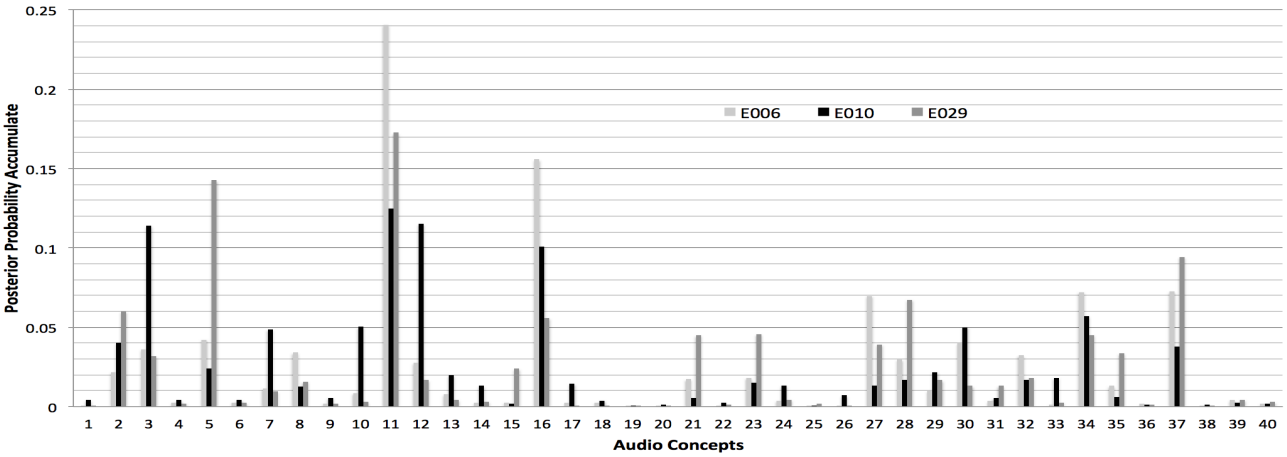

Figure 2: Probability histogram of the 40 audio concepts accumulated across three sample events, based on the 100 training video files. E006 = *Birthday party*; E010 = *Grooming an animal*; E029 = *Winning a race without a vehicle*. Each event shows higher detection probabilities for different concepts.

Label	Event Description	Cnd.1: MFCC	Cnd.2: H-DNN	Cnd.3: MFCC+ H-DNN
E006	Birthday party	36	**44**	43
E007	Changing a vehicle tire	**12**	5	11
E008	Flashmob gathering	**45**	21	32
E009	Getting a vehicle unstuck	22	**27**	17
E010	Grooming an animal	**13**	11	7
E011	Making a sandwich	**20**	11	17
E012	Parade	23	**35**	28
E013	Parkour	39	42	**51**
E014	Repairing an appliance	32	**58**	44
E015	Working on a sewing project	15	10	**17**
E021	Attempting a bike trick	27	**40**	33
E022	Cleaning an appliance	**28**	12	**28**
E023	Dog show	**27**	**27**	**27**
E024	Giving directions to a location	9	**19**	13
E025	Marriage proposal	12	**15**	**15**
E026	Renovating a home	**22**	9	19
E027	Rock climbing	**28**	17	6
E028	Town hall meeting	**63**	53	47
E029	Winning a race without a vehicle	18	**23**	**23**
E030	Working on a metal crafts project	**14**	5	9
ALL	Mean Avg. Precision	26.03	26.37	26.71

Table 2: Percentage precision performance for each experimental condition, per event and overall. Each feature type generally excels at detecting different events, as can be seen for each of the three conditions.

As we noted above, the vocabulary of audio concepts we are using includes some with very similar labels. However, the concepts are not detected in an equal degree for the videos they would seem at first glance to correspond to. For instance, *singing SRI* is detected much more frequently than *singing CMU*, which has a low cumulative probability across the events; similarly for *birds STAND* vs. *anim bird CMU*—though both had several times stronger correspondences with *Grooming an animal* than with the other types of events. This discrepancy confirms that, even when the labels used are very similar, we cannot assume that the concepts should be merged.

The vocabulary of 40 sounds used in these experiments is relatively small for distinguishing 20 different video events, thus some events are more distinguishable from audio-concept evidence than others. A larger vocabulary, containing more event-relevant audio concepts, would most likely improve discrimination. Our system outputs a posterior probability for at least one of the 40 audio concepts for each frame regardless of whether the frame actually matches any of the concepts.

4.2. Video Event Detection Performance

For comparison purposes, the HMM-based MED system is fed with three different types of features:

Condition 1 (Baseline): MFCCs of 13 + Δ + $\Delta\Delta$ coefficients (dimensionality 39);

Condition 2: H-DNNs of 40 audio-concepts described in Section 3.1;

Condition 3: A tandem combination of the MFCCs 13 coefficients and the the H-DNN features (dimensionality 53).

For each different feature type, a different HMM architecture (i.e. the number of gaussian mixtures, hidden states and transition probabilities) is derived by optimizing the video-detection performance over a small development set (10% of the total training data). In sum, we employed 11 states with 128 diagonal gaussians each for MFCCs (Condition 1), 9 states with 64 Gaussians each for H-DNN features (Condition 2), and 9 states with 128 gaussians for the tandem MFCC+H-DNN features (Condition 3).

For training each HMM model we used the EK100 set, consisting of 100 videos each for 20 event categories (listed

in Table 2), a total of 2,000 video files. For testing, we used all the positive video files, about 1500, from the MED test set, i.e., those positively categorized as belonging to one of the 20 events. As large numbers of non-class or negative videos have been shown to significantly affect performance on event detection, we analyzed only videos that belong to an event, to obtain more reliable conclusions. Thus, we left out the approximately 5,000 negative training files and approximately 23,000 negative testing files.

Detection performance per event for the three types of features is shown in Table 2. The mean average precision (MAP) per event does not differ significantly across the three conditions. The MAP was for MFCCs 26.03%, for H-DNNs 26.37%, and for MFCC+H-DNNs 26.71%. However, one of the most important results from these experiments is that they show that, even with a limited set of 40 audio concepts, the discriminative power of the proposed system was competitive with the system using MFCCs.

The range of the precision scores across events shows more detailed differences between the different types of features. The MFCC condition (1) has the widest range: between E024 *Giving directions to a location* at 9.4% and E028 *Town Hall Meeting* at 63.2%, a spread of 53.8%. The H-DNNs (Condition 2) have a narrower range: between E030 *Working on a metal crafts project* at 4.5% and E014 *Repairing an appliance* at 52.6%, a spread of 48.1%. Finally, the MFCC+H-DNN combination (Condition 3) has the most consistent precision across events: between E027 *Rock climbing* at 5.6% and E013 *Parkour* at 50.5%, still a spread of 44.9%.

Confusion between events seems to differ depending on which types of features are used for analysis, further confirming that each feature type best addresses different types of acoustic information. For example, E006 *Birthday Party* is most often confused with E008 and E028 using MFCCs; with E029 and E013 using H-DNNs; and with E013 and E011 using for MFCCs+H-DNNs. The MFCC-only results have the most individual events detected with the highest precision, with 8 of the highest per-event scores, but it has the lowest MAP of the three conditions. Conversely, the MFCC+H-DNN results have the fewest events detected with the highest precision, only 2, but the overall MAP score is the highest. The H-DNN-only results are in the middle, with 6 of the highest-precision events. In other words, if the goal is to achieve high performance on specific events, then it is better to use either MFCCs or H-DNNs, whichever is better for that event. However, if the goal is to achieve balanced performance across events, MFCCs+H-DNNs is the best choice. Due to space contraints, we show only the confusion matrix for the MFCC+H-DNN combination Fig. 3

There are four event types, E008 (**45%**,21%,32%), E014 (32%,**58%**,44%), E027 (**28%**,17%,6%), and E028 (**63%**,53%,47%), where one feature type outperforms the others by at least an absolute 10%. This suggests that the different types of features address different acoustic characteristics. Specifically, MFCCs seem to be stronger at characterizing speech-related events, while the H-DNNs are better for characterizing events with mixed frequency and wider-spectrum sounds.

It is worth noting that, for E023 *Dog Show*, the three conditions yield equal precision; when we spot-checked example videos, we found a prominent white noise-like sound common in recordings of crowded events, as well as background speech, and barking; less frequently, there is also music. Except from barking, none of them seem to be different from other public events, which is why the event yields lower-than-average precision. Moreover, it is interesting that combining MFCCs and H-DNNs does not provide a significant gain for most events. It is possible that the combination could benefit from a dimensionality reduction technique. More insights on how to maximize performance of the fusion should emerge from future work.

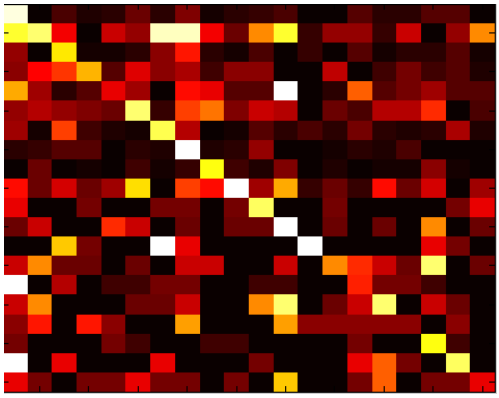

Figure 3: Confusion matrix for Condition 3, the MFCC+H-DNNs-based video-detection system (MAP 26.71%).

5. Conclusions and Future Work

This paper describes a multimedia event detection system that exploits the temporal correlation of audio concepts in an event. First, the audio-concept features are extracted using a Hierarchical Deep Neural Network, which analyzes the short- and long-term surrounding acoustic information for the concept. Second, Hidden Markov Models are used to model the continuous, non-stationary characteristics of the audio signal throughout the event; the HMMs are then used for event detection in videos. The resulting system performs competitively in comparison with an MFCC-based system, but because it involves audio concepts, it can show humanly understandable evidence to explain the relationship of those audio concepts with events in video. Although the ratio of audio concepts to event types minimizes the possibilities for discrimination, the system can take advantage of even this limited selection of audio concepts to distinguish some events with fair reliability. Certainly, the next step for the research presented here is to include the negative files from the test set and analyze their impact. In addition, it is important to expand the set of audio concepts.

6. Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior National Business Center contract number D11PC20066. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsement, either expressed or implied, of IARPA, DOI/NBC, or the U.S. government. Also supported by Lawrence Livermore National Laboratory, which is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration, under Contract DE-AC52-07NA27344.

7. References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, Eds. 2012.

[2] Shoou-I Yu, Zhongwen Xu, Duo Ding, Waito Sze, Francisco Vicente, Zhenzhong Lan, Yang Cai, Shourabh Rawat, Peter Schulam, Nisarga Markandaiah, Sohail Bahmani, Antonio Juarez, Wei Tong, Yi Yang, Susanne Burger, Florian Metze, Rita Singh, Bhiksha Raj, Richard Stern, Teruko Mitamura, Eric Nyberg, and Alexander Hauptmann, "Informedia e-lamp @ TRECVID 2012, multimedia event detection and recounting," 2012.

[3] Pradeep Natarajan, Prem Natarajan, Shuang Wu, Xiaodan Zhuang, Amelio Vazquez-reina, Shiv N. Vitaladevuni, Carl Andersen, Rohit Prasad, Guangnan Ye, Dong Liu, Shih fu Chang, Imran Saleemi, Mubarak Shah, Yue Ng, Yn White, Larry Davis, Abhinav Gupta, and Ismail Haritaoglu, "BBN VISER TRECVID 2012 multimedia event detection and multimedia event recounting systems," 2012.

[4] Hui Cheng, Jingen Liu, Saad Ali, Omar Javed, Qian Yu, Amir Tamrakar, Ajay Divakaran, Harpreet S. Sawhney, R. Manmatha, James Allan, Alex Hauptmann, Mubarak Shah, Subhabrata Bhattacharya, Afshin Dehghan, Gerald Friedland, Benjamin Martinez Elizalde, Trevor Darrell, Michael Witbrock, and Jon Curtis, "SRI-Sarnoff AURORA system at TRECVID 2012, multimedia event detection and recounting," 2012.

[5] Xiaodan Zhuang, Stavros Tsakalidis, Shuang Wu, Pradeep Natarajan, Rohit Prasad, and Prem Natarajan, "Compact audio representation for event detection in consumer media," in *INTERSPEECH*, 2012.

[6] Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange, and Mark D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY*, 2013.

[7] Qin Jin, Peter F. Schulam, Shourabh Rawat, Susanne Burger, Duo Ding, and Florian Metze, "Event-based Video Retrieval Using Audio," in *Proceeding of the 13th Annual Conference of the International Speech Communication Association*, 2012.

[8] Stephanie Pancoast, Murat Akbacak, and Michelle Sanchez, "Supervised Acoustic Concept Extraction for Multimedia Event Detection," in *ACM International Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis at ACM Multimedia*, 2012.

[9] Anurag Kumar, Pranay Dighe, Rita Singh, Sourish Chaudhuri, and Bhiksha Raj, "Audio Event Detection from Acoustic Unit Occurrence Patterns," in *ICASSP*, 2012.

[10] Feng Su, Li Yang, Tong Lu, and Gongyou Wang, "Environmental sound classification for scene recognition using local discriminant bases and HMM.," in *ACM Multimedia*. 2011, ACM.

[11] Diego Castan and Murat Akbacak, "Segmental-gmm approach based on acoustic concept segmentation," in *SLAM@INTERSPEECH*, 2013.

[12] Diego Castan and Murat Akbacak, "Indexing Multimedia Documents with Acoustic Concept Recognition Lattices," in *Interspeech 2013*, 2013, pp. 3–7.

[13] Benjamin Elizalde, Mirco Ravanelli, and Gerald Friedland, "Audio concept ranking for video event detection on user-generated content.," in *SLAM@INTERSPEECH*, 2013.

[14] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Wessel Kraaij, Alan F. Smeaton, and Georges Quénot, "TRECVID 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2013*. NIST, USA, 2013.

[15] Susanne Burger, Qin Jin, Peter F. Schulam, and Florian Metze, " Noisemes: Manual Annotation of Environmental Noise in Audio Streams ," Tech. Rep., 2012.

[16] Mirco Ravanelli, Benjamin Elizalde, Karl Ni, and Gerald Friedland, "Audio concept classification with hierarchical deep neural networks," in *European Signal Processing Conference*, 2014.

[17] Mirco Ravanelli, Van Hai Do, and Adam Janin, "TANDEM-Bottleneck Feature Combination using Hierarchical Deep Neural Networks," in *ISCSLP 2014*.

[18] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[19] Karel Vesely, Lukas Burget, and Frantisek Grezl, "Parallel Training of Neural Networks for Speech Recognition," in *Proceedings of INTERSPEECH*, 2010.

[20] Lawrence R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, 1989, pp. 257–286.

[21] Steve J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*, Cambridge University Press, 2006.