Event-based Video Retrieval Using Audio

Qin Jin, Peter F. Schulam, Shourabh Rawat, Susanne Burger, Duo Ding, and Florian Metze

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

{qjin,pschulam,srawat,sburger,dding,fmetze}@cs.cmu.edu

Abstract

Multimedia Event Detection (MED) is an annual task in the NIST TRECVID evaluation, and requires participants to build indexing and retrieval systems for locating videos in which certain predefined events are shown. Typical systems focus heavily on the use of visual data. Audio data, however, also contains rich information that can be effectively used for video retrieval, and MED could benefit from the attention of researchers in audio analysis. We present several systems for performing MED using only audio data, report the results of each system on the TRECVID MED 2011 development dataset, and compare the strengths and weaknesses of each approach.

Index Terms: multimedia event detection, audio processing, video categorization

1. Introduction

Growing amounts of multimedia data available on the Internet are making the development and advancement of audio and video information retrieval technologies invaluable [1]. Multimedia Event Detection (MED) is the problem of indexing and searching large corpora of video data in order to retrieve from the collection those videos that show instances of certain predefined events. Examples of such events in past MED evaluations have been “constructing a shelter” and “baking a cake,” and are defined using event kits, which are textual descriptions of the desired event along with several example videos. Traditional multimedia information extraction techniques have focused largely on the visual domain. Audio, however, can also contribute significant information when searching for particular events. Identifying spoken utterances can provide linguistic evidence, while certain classes of environmental sounds can suggest that a particular activity or object is being shown in the video. Audio is useful especially in situations when other sensors such as video fail to reliably detect the events. For example, when visual evidence is distorted or unreliable due to poor lighting conditions, microphones can pick up acoustic evidence that can complement or enhance the limited visual data available. When both audio and video data can be reliably extracted and analyzed, systems can make detection decisions with higher confidence [2]. With these motivations in mind, we present five different audio-only multimedia event detection systems, along with experimental results on the TRECVID MED 2011 development data. Our primary contribution is a comparative evaluation of a diverse set of approaches to audio-based MED.

2. Related Work

While there has been interest in acoustic scene analysis and audio processing for multimedia databases since the nineties, there have been significant shifts in the types of data from which researchers have been focused on retrieving information. Generally, the setting in which one wishes to extract information from soundtracks can be classified according to three general dimensions.

First, the quality of the audio data can vary considerably. Early work on audio event classification was largely done on sound databases [3] and clean broadcast or television program audio data [4]. It is generally easy to distinguish important foreground data from the background noise in typical high quality database or broadcast recordings. Multimedia audio conditions are rarely ideal, and present considerable challenges for audio MED.

Second, the data can differ with respect to the number of sounds that one wishes to identify within the audio. Much work in the nineties was focused on discriminating only between speech and music, and was solved using a wide array of traditional machine learning and signal processing approaches [4, 5, 6]. As is well understood in machine learning, increasing the number of classification categories typically makes a task more difficult.

Finally, the task can differ with respect to the granularity of the audio processing that one wishes to perform. The aforementioned efforts to segment speech and music in broadcasts are sub-soundtrack problems in which the goal is to produce transcriptions of the audio according to some fixed set of labels. Alternatively, we can classify the entire soundtrack. For example, we may wish to determine the environment in which a particular soundtrack was recorded [7, 8]. Such tasks have become especially relevant in today’s world, where portable, personal devices are prevalent and capable of recording and storing large amounts of audio and video data. In this paper, our focus is similar to [9], in which the goal is to detect whether certain events or activities are occurring in a video by relying only on audio evidence. Only for the segmenter described in Section 4.3 do we perform sub-soundtrack transcription before classifying the entire video.

3. Data and Experimental Setup

Our experiments within this paper are conducted on the development data from the TRECVID 2011 Multimedia Event Detection (MED) [10] task. Fifteen events are defined: Attempting a board trick (E001), Feeding an animal (E002), Landing a fish (E003), Wedding ceremony (E004), Working on a woodworking project (E005), Birthday party (E006), Changing a vehicle tire (E007), Flash mob gathering (E008), Getting a vehicle unstuck (E009), Grooming an animal (E010), Making a sandwich (E011), Parade (E012), Parkour (E013), Repairing an appliance (E014), Working on a sewing project (E015).

The dataset contains 3104 videos for training, and 6642 videos for testing. Within the training set, each event has around 90 positive video samples. The remaining videos are unrelated to the target events listed above, and are provided as “background” videos.

In addition to the disjoint training and testing sets, we manually labeled environmental noise for the soundtracks of 216 videos taken from the training set (totaling 5.6 hours). We picked at least 10 videos for each event, added videos from three more events used in MED 2010 (batting in a run, making a cake and building a shelter), and 26 more “background” videos. The audio tracks were annotated using the PRAAT tool. We identified and segmented distinct kinds of noise in the audio stream (by listening to the noise as well as observing the signal in both the time and frequency domain) and described the noise segments in a detailed open labeling approach. We sorted, combined and categorized the open labels according to their spectrogram and temporal properties (formant, pulse-like, friction, fuzziness) and their possible sources. The result is a set of 42 noise units, which we call "noisemes" [11]. The original video/image data was not part of the annotation to avoid the influence of visual concepts whenever possible: e.g. there is no noiseme for car in our set. There can be a multitude of sounds associated with a car, some of which are only identifiable by observing a car in the video. Our goal was to create labels for audio segments that “exist” solely in the audio domain.

The evaluation criteria for the event detection system performance include DET curve, average Pmiss@TER=12.5% and average minNDC [10]. In Figures 2, 3 and 4 we provide a side-by-side comparison of three of the best systems. The key trends to note are the ways in which the DET curves for each system move towards or away from the lower left corner.

4. Systems Description

We now present the five audio-based MED systems that we have developed. Figure 1 illustrates the general system components of an audio event detection system. Each system presented in the following sections shares this particular architecture. The differences are in the types of features extracted from the audio data, the models used for representing each of the 15 events, and the methods used for training the models.

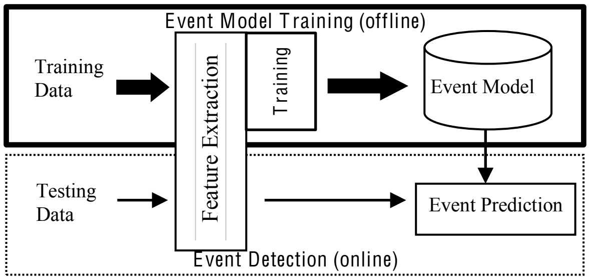

Figure 1: *Audio Event Detection System.*

4.1. GMM Super Vector with SVM (GSV-SVM)

Support vector machines trained on Gaussian “supervectors” have been successfully used on the speaker verification task [12]. A GMM supervector (GSV) is constructed by stacking the means, diagonal covariance vectors, and/or component weights of the mixture model. A support vector machine operating in the supervector feature space is then used to classify clips of speech. We explored building an SVM classifier based on Gaussian supervectors for audio event detection. The Gaussians are used to model 40 dimensional MFCCs (20 coefficients, and 20 deltas) extracted from a video’s soundtrack, and use diagonal covariance matrices. We first trained a universal background model (UBM) by sampling audio from a set of videos that do not belong to our training or testing data. We then created vector representations of testing and training videos using MAP adaptation to adjust the means and covariances of each of the component Gaussians given the MFCC feature vectors extracted from the video. To create a super vector from the mixture of Gaussians, we concatenate the diagonals of each of the Gaussian covariance matrices into a single vector. When added to the super vector, we take the square root of the diagonal values, and multiply each value by the square root of the weight of that Gaussian in the mixture model. We then use the supervector as a feature vector to train the SVM event classifier. The GSV-SVM system achieves average Pmiss@ TER=12.5% of 0.571 and minNDC of 0.855 using 4096 components in the UBM GMM (Figure 2).

4.2. Data-driven MFCC based Bag-of-Words

Another data-driven approach is to define the audio concepts in terms of a codebook or a bag-of-audio-words model. The codebook model is a common technique used in document classification (bag-of-words) and [15] image classification (bag-of-visual words). The bag-of-audio-words has also been applied to the MED tasks [13]. The bag-of-audio-words model represents an audio file by quantizing low level audio features into a discrete set of code words in the “vocabulary” (codebook) thus providing a histogram of codeword’s counts. These code words are either learned through manual annotation or via unsupervised clustering. The discriminative power of such a codebook is governed by the size of the codebook and by the assignment of features to code words [14]. In this paper we apply this model to the task of Multimedia Event Detection using MFCC features. We use 40-dimensional feature vectors (20 MFCCs + 20 deltas). MFCCs are computed every 32ms with 50% overlap (16ms shift). The vocabulary is learned by randomly sampling 3 million samples from the whole dataset and clustering into 16,000 codewords using k-means. Each video is then represented as a distribution over these 16000 codewords by using soft-assignment of MFCC features to these codewords. We then train a one-against-all SVM classifier using the $\chi 2$ kernel for each of the 15 events over these bag-of-audio-words features. The MFCC Bag-of-Audio-Words Model achieves an average PMiss@TER=12.5% of 0.558 and minNDC of 0.846 over the 15 events.

4.3. Bag-of-Words with Audio Segmentation

Many approaches to multimedia event detection using audio features focus on identifying the setting or environment of the entire video given the acoustic features extracted from the soundtrack (e.g. MFCCs). Lee and Ellis [9] predict the type or content of a video using single Gaussian modeling of the MFCCs, Gaussian mixture modeling, and latent semantic analysis of

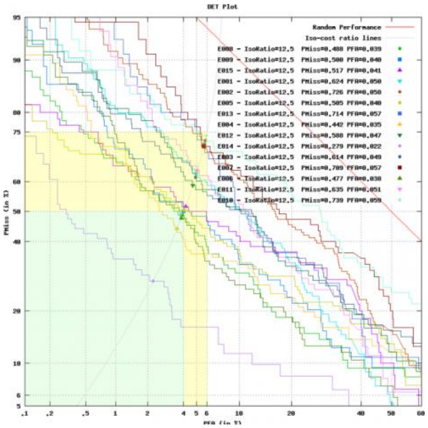

Figure 2: DET Curve for GSV-SVM

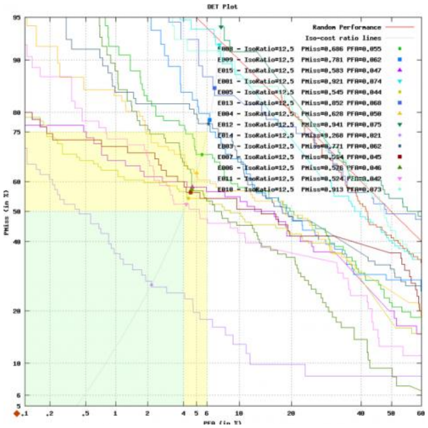

Figure 3: DET curve for ASR

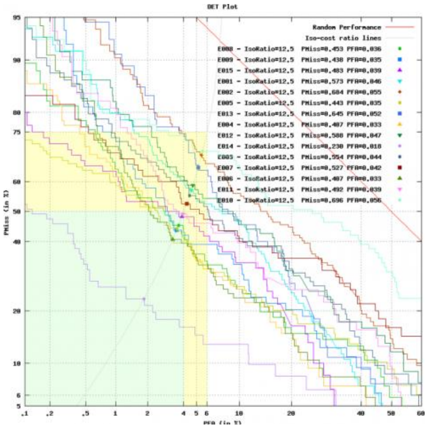

Figure 4: DET Curve for Fused System

Gaussian component histograms. It would, however, be useful for multimedia information extraction applications to have access not only to the type of the video's soundtrack, but also the specific types and temporal locations of sounds heard within the soundtrack.

For our audio-based multimedia event detection system we experimented with a more fine-grained approach to audio event analysis based on techniques borrowed from traditional automatic speech recognition systems. This approach can be broken down into two distinct steps. First, we use a hidden Markov model audio event segmenter that assigns labels to intervals of the soundtrack. Labels are chosen from the set of 42 noisemes. Second, we use a bag-of-words feature vector computed from summing the total occurrences of each type of sound label for a particular soundtrack, and use this bag-of-words feature vector as input to a traditional one-against-all SVM classifier. We build such a classifier for each MED target event.

The segmenter is based on traditional continuous density hidden Markov model automatic speech recognition systems. Our set of "phonemes" is simply the set of audio event labels. The dictionary used for decoding contains one "word" for each audio event composed entirely of that audio event's phoneme. Nearly all audio event dictionary entries are composed of a single phoneme. The only exceptions are audio events that tend to occur over longer intervals such as speech, music and silence. All hidden Markov models for each phoneme are tri-state, sequential HMMs. Unlike traditional tri-state HMMs used in automatic speech recognition, each state in the audio event HMMs are identical (as opposed to beginning, middle, and end states). The length of each HMM simply imposes a minimum frame length for our audio events.

We achieve a 51.96% overall speaker diarization error computed using 10-fold cross validation over the 216 videos for which we have ground truth annotations. Introducing a simple "language model" into our segmenter results in a 0.5% absolute improvement; dropping the overall speaker diarization error to 51.47%. With respect to MED, the segmenter achieved an average Pmiss @ TER=12.5% of 0.746 and minNDC of 0.948.

4.4. Phonetic Bag-of-Words

The phonetic/multilingual phone stream approach has been successfully used in the speaker recognition task [13]. We applied it on the audio event detection task. The general idea is that we can represent each audio with a sequence of phone tokens. And then we can use bag-of-words features on the phone tokens to train event classifiers. Each audio can be represented by multiple phone sequences and therefore we can build multiple event classifiers for event detection and combine them. We started to tokenize the audio files with 13 open-loop phone recognizers which were trained on the Global Phone corpus, including Arabic, Mandarin, German, English, French, Japanese, Korean, Croatian, Portuguese, Russian, Spanish, Swedish, and Turkish. We then studied how to combine and fuse multiple single systems including early fusion of multiple phone streams and late fusion of detection scores etc. By early fusion, we pooled the phone streams from multiple phone recognizers together and extracted the bag of words features on the merged phone stream. By late fusion, we simply linearly fused the event detection scores from different single phonetic systems with equal weights. The early fusion (Pmiss@TER=12.5% of 0.631 and minNDC of 0.894) achieved slightly better performance than the late fusion (Pmiss@TER=12.5% of 0.635 and minNDC of 0.898).

4.5. ASR Bag-of-Words

Expressing similarities between clips using the hypotheses of an automatic speech recognition (ASR) system has been an established basis for performing video retrieval for quite some time. On the present database, only about 60% of videos in the development set do contain speech at all, with a significant overlap with noises, and generally difficult conditions.

To perform ASR, we perform an initial segmentation pass using an ergodic HMM with speech, silence, music, and noise states, trained on BN-style data, and essentially a simple version of the "segmenter" described above. We then perform speech-to-text on the "speech" parts, using a wide-band MFCC acoustic model EM-trained on a mixture of BN and "Meeting" training data, and a conversational language model, with a vocabulary of around 50k. No adaptation is performed, and we measure a word error rate of around 60% on the development set, with a significant amount of deletions (around 20%) and substitutions (ca. 35%), with only few insertions. This configuration gave best retrieval performance, and appears to be robust across a number of conditions observed in the data.

For classification, we use standard Porter stemming, followed by a bag-of-words classifier, which has been trained on the development set. The performance is shown in Figure 3

(Pmiss@TER=12.5% of 0.701 and minNDC of 0.892). Event-specific Pmiss values differ more for ASR, than for the other approaches, given that only about 60% of videos contain speech, and this distribution correlates with events. E014 (repairing an appliance) performs best for ASR and GSV-SVM. E007 (changing a vehicle tire) performs very well for ASR, while it does not perform as well for GSV-SVM, so approaches are complementary. As we would expect, ASR-based performance is better for events that, on average, contain more speech, such as "How To"-style events (e.g. repairing an appliance or making a sandwich).

4.6. Fusion of Multiple Systems

We simply fused the multiple systems with equal weights. The fused system performance is shown in Figure 4. Table 1 summarizes the performance of the 6 individual audio detection systems as described above and the fused system. Fusion of individual systems significantly improves the performance. More advanced fusion strategies will be explored in our future work.

Table 1. Performance of sub-systems and fused system

Systems	Pmiss@TER=12.5%	minNDC
GSV-SVM	0.571	0.855
MFCC K-means	0.558	0.846
Segmenter	0.746	0.948
Phonetic early Fusion	0.631	0.893
Phonetic late fusion	0.635	0.898
ASR	0.701	0.892
Fusion	0.510	0.798

5. Discussion and Conclusions

We have presented five different multimedia event detection systems based entirely on audio-based features. It is not surprising that the data driven systems (GSV-SVM, MFCC K-means) achieve the best overall individual performances on our dataset. Both systems are based on unsupervised modeling techniques, and are capable of picking up subtleties in the data. Phonetic early and late fusion show promising results, but can ultimately be replaced by the GSV-SVM and MFCC K-means system given that they are neither as successful as the purely data-driven approaches, and not as interpretable as the ASR and segmenter approaches. Finally, the ASR and segmenter systems were not generally effective by themselves (although ASR performs as well if not better than data-driven systems given the presence of speech). We believe, however, that while the ASR and segmenter approaches may not be best for pure MED, both can be used effectively in tasks where detailed, user-interpretable summaries of videos are required as output. For such tasks, it is important to have information regarding the presence, duration, and relative order of spoken utterances and certain types of audio events. It is interesting to note that fusing the output of each of the systems into a single classifier yields significant improvements for both Pmiss@ TER=12.5% and minNDC. Furthermore, we fused systems simply by combining outputs, more elaborate techniques may yield better results.

There are two directions in which we will direct future research on our audio-based MED systems. First, we would like to improve the accuracy of our data-driven systems by introducing more structure into the classification process. We currently treat the entire video as a large chunk of data, and ignore temporal and durational aspects of the features that occur within the soundtrack.

Second, we would like to improve both the segmenter and ASR systems. The primary limiting factor for these two systems is the availability of training data. We are currently exploring ways to utilize the massive amounts of unlabeled audio data available to us in conjunction with our small set of labels to make the segmenter and ASR system more robust.

6. Acknowledgements

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

7. References

[1] Naphade, M.R. and Huang, T.S., “Multimedia understanding: challenges in the new millennium”, International Conference on Image Processing, 2000, pp 33-37.

[2] Atrey, P.K., Kankanhalli, M.S. and Jain, R., “Information Assimilation Framework for Event Detection in Multimedia Surveillance Systems”, Multimedia Systems, 2006, pp 239-253.

[3] Wold, E., Blum, T., Keislar, D., and Wheaten, J., “Content-based Classification, Search, and Retrieval of Audio”, IEEE Multimedia, 3(3):27-36, 1996.

[4] Saunders, J., “Real-time Discrimination of Broadcast Speech/Music”, Proc. of IEEE Intern. Conference on Acoustics, Speech and Signal Processing, 2:993-996, 1996.

[5] Scheirer, E. and Slaney, M., “Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator”, IEEE Conf. on Acoustics, Speech, and Signal Processing, 2:1331-1334, 1997.

[6] Williams, G. and Ellis, D.P.W., “Speech/Music Discrimination Based on Posterior Probability Features”, Proc. Eurospeech, 1999.

[7] Ma, L., Milner, B., and Smith, D., “Acoustic Environment Classification”, ACM Transactions on Speech and Language Processing, 3(2):1-22, 2006.

[8] Eronen, A., Peltonen, V., Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T., Lorho, G., and Huopaniemi, J., “Audio-based Context Recognition”, IEEE Trans. on Audio, Speech, and Language Processing, 14(1):321-329, 2006.

[9] Lee, K. and Ellis, D.P.W. “Audio-Based Semantic Concept Classification for Consumer Video”, IEEE Trans. Audio, Speech, and Language Processing, 18(6):1406-1416, 2010.

[10] TRECVID: http://www-nlpir.nist.gov/projects/tv2011/tv2011.html

[11] Burger, S., Jin, Q., Schulam P.F., and Metze, F., “Noisemes: Manual Annotation of Environmental Noise in Audio Streams”, in submission to Interspeech 2012.

[12] Campbell, W.M., Sturim , D.E. and Reynolds, D.A. “Support vector machines using GMM supervectors for speaker verification”, IEEE Signal Processing Letters, 2006, pp 308-311.

[13] Jin, Q., Schultz, T. and Waibel, A., “Phonetic Speaker Identification”, International Conference of Spoken Language Processing (ICSLP-2002), 2002.