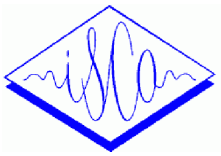
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Robust Event Detection From Spoken Content In Consumer Domain Videos

Stavros Tsakalidis, Xiaodan Zhuang, Roger Hsiao
Shuang Wu, Pradeep Natarajan, Rohit Prasad, Prem Natarajan

Raytheon BBN Technologies
10 Moulton St., Cambridge, MA 02138
{stavros, xzhuang, whsiao, swu, pradeepn, rprasad, pnataraj}@bbn.com

10.21437/Interspeech.2012-560

Abstract

In this paper, we propose an innovative integrated approach to leverage available spoken content while detecting events in consumer-generated multimedia data (i.e., YouTube videos). Spoken content in consumer videos exhibits several challenges. For example, unlike Broadcast News, the spoken audio is typically not labeled. Also, the audio track in consumer videos tends to be noisy and the spoken content is often sporadic.

Here, we describe three recent improvements that are specifically targeted at overcoming the challenges in consumer videos: robust data-driven keyword selection, automatic discovery of word-classes for keyword expansion, and a keyword spotting approach for improving recall in noisy conditions. These improvements are integrated into the audio analysis component of the BBN VISER system. The VISER system embodies a state-of-the-art approach as substantiated by its performance on the 2011 TRECVID MED task. Experimental results on the 2011 TRECVID MED task clearly demonstrate the effectiveness of the three improvements.

Index Terms: multimedia event detection, keyword selection, keyword expansion, keyword spotting.

1. Introduction

The ability to automatically search through large volumes of digital videos and summarize their content has several applications including retrieval and copy detection. Until recently, the main focus of research in digital video analysis and understanding has been on professional sources such as broadcast news organizations. Nowadays, the abundance of inexpensive video recording devices, such as the ones found in most smart phones, and the simplicity of uploading and sharing the video recordings online has resulted in a dramatic explosion of consumer domain videos. The ever expanding volume of consumer domain videos on the Internet has triggered an interest in automatically analyzing such content.

Most web videos contain a complementary stream of information in the accompanying audio track. Audio has lower dimensionality compared to the video track, and it is relatively invariant with respect to video quality and viewpoint. Therefore, audio provides a powerful mechanism for video query and retrieval, as well as for the recognition of complex actions that may be difficult, or even impossible to perform using video alone. Furthermore, human language is often present in consumer videos in the form of spoken content in the audio track. Such content could potentially provide useful information for detecting events of interest. Besides, the semantic information from human language is, typically, complementary to the information from low-level audio and visual features.

The development of technologies for spoken content analysis found in web videos faces many challenges. Web videos are highly heterogeneous due to variability in content, style, production qualities and language. The speech is typically spontaneous, conversational and characterized by high variability and lack of inherent structure. Background noise and speech from multiple background speakers overlaid on the acoustic signal of the target speaker are often encountered in web videos. Also, due to the large volume of such heterogeneous data, techniques that rely on manually labeled data are impractical. Last but not least, the approach should be scalable and able to handle a large, or varying, set of target events in a huge archive, requiring minimum human intervention.

In this paper, we propose a comprehensive solution to the event detection problem in multimedia data by leveraging the speech content. We present recent improvements in the spoken content analysis component of the BBN VISER system [1]. BBN's VISER system embodies a state-of-the-art approach as substantiated by its performance on the the Multimedia Event Detection (MED) task of the 2011 TREC Video Retrieval Evaluation (TRECVID).

The cornerstone of our approach is to use several data-driven, automated learning and unsupervised techniques. Specifically, we improve the baseline VISER system by using robust procedures that address the absence of labeled speech audio content, noisy audio conditions, and sparsity of data. First, we improve the data-driven keyword selection from automatically transcribed and therefore unreliable transcripts by using word-level confidence measures. We then employ a keyword spotting approach instead of a speech-to-text transcription approach to improve keyword discovery in noisy conditions. We also propose a score mapping technique which can learn a mapping between the arc posterior probability of a keyword in the recognition lattice to the probability of being correct. Finally, in order to train with limited training data, we adopt a keyword expansion approach via automatic discovery of semantically related word-classes.

2. Robust Data-Driven Keyword Selection

To leverage the hypothesized speech transcripts in multimedia event detection, we use the distribution of event-discriminating

Acknowledgement: Supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20071. 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 endorsements, either expressed or implied , of IARPA, DoI/NBC, or the U.S. Government.

keywords within each video clip. This bag-of-words feature representation is used to train kernel-based classifiers for each event. The event-discriminating keywords are automatically discovered via the Term Frequency Inverse Document Frequency (TF-IDF) metric [2]. TF-IDF selects features based on how relevant a given word is in a particular class.

TF-IDF has been used extensively for document retrieval. However, their efficacy diminishes when dealing with unreliable labels such as the output of an automatic speech recognizer (ASR). Thus, it is essential to automatically assess the reliability of the automatically recognized words. Confidence measures enable us to assess the output of an ASR system [3]. The confidence measure provides us with the posterior probability that a hypothesized word is correct. We use a modified version of TF-IDF [4] to exploit word confidences in order to reduce the inherent uncertainty in the ASR output.

Starting with the hypothesized speech transcripts, we remove stop words, and then normalize them using the Porter stemmer [5]. Then, for each event we choose the top-N words according to the following modified TF-IDF metric:

$$\left(\frac{n}{t}\right)\log\left(\frac{d}{h}\right) \tag{1}$$

where n is the confidence-weighted number of times a word appears in video clips belonging to a particular event category; t is the total confidence-weighted number of words in that category; d is the total number of categories considered; h is the number of categories containing the word. Finally, we take the union of the event-dependent keyword lists to derive the final keyword list. This approach can be extended by considering multiword keywords.

3. Word Clusters for Keyword Expansion

Though the automatically identified keywords can distinguish between events, their distributions within the events are sporadic. A low number of training examples for an event can lead to low recall of similar unseen events due to the sparsity of the keywords. This behavior is greatly exaggerated when we deal with generic events, such as "making a sandwich", and highly heterogeneous video collections. One way to make up for lack of data is to substitute some form of knowledge for it. A commonly used approach to overcome this issue is query expansion [6], where videos are retrieved not only with the initial query keywords query related to the event, but also with closely related terms.

We can leverage an automatically generated thesaurus to alleviate the sparseness of keywords and automatically induce the relationship between the words. We refer to sparseness here as the phenomenon that the informative keywords in different video clips often do not overlap, though they carry highly related semantic information, owing to the limited availability of speech content and imperfect ASR output in this domain. To bridge such a gap between semantically related words, we extent the keyword list with automatically-derived class-based features. First, we adopt an external thesaurus with automatically derived proximity-based similarities for many common word pairs [7]. The similarity is computed based on the linear proximity relationship between words. Second, based on such pairwise similarities, we employ the affinity propagation clustering algorithm [8] which automatically infers a set of mutually exclusive word classes for the keywords. The algorithm identifies exemplars among data points and forms clusters of data points around these exemplars. It operates by simultaneously considering all data points as potential exemplars and exchanging messages between data points until a good set of exemplars and clusters emerges.

4. Confusion Networks for Event Detection

Using the 1-best output of the recognizer may not be optimal for event detection. Noise, channel mismatch, domain mismatch may degrade the recognition performance and lead to high Word Error Rate (WER). High WER implies that the keywords in the audio may not be recognized, and as a result, the system may fail to detect an event. To tackle this problem, one can use a lattice or confusion network [9] instead of 1-best for keyword spotting [4]. Compared to 1-best, a lattice contains more alternatives and it is more likely to recover some of the keywords in the lattice. However, some keywords in the lattices may actually be false alarms. Hence, we need some kind of scoring methods to evaluate whether a word in a lattice is an actual hit or a false alarm.

One of the simplest scoring methods is using the arc posterior probabilities derived via the forward-backward algorithm [10]. If a word in the lattice has high posterior probability, the recognizer firmly believes such word exists in the audio, otherwise, the word would have a lower posterior probability, and it is more likely to be a false alarm. Posterior probability of a word can be served as a soft count for that word. Given an audio clip, one can collect the soft counts from the lattices and create a histogram to represent the audio clip [4]. Such a histogram is used for classification afterwards. To avoid the noise created by the words with very low posterior probability, we can use a threshold to remove the words with posterior probability lower than the threshold. Histogram can also be constructed using only the 1-best in a similar fashion where every word in the 1-best is counted as one occurrence, and one can collect the counts to create the histogram.

While posterior probability may help to determine whether a word is a true hit or a false alarm, the scale may be affected by how one tunes the recognizer or the lattice generation process. For example, a word with posterior probability in the range of 0 to 0.70 may have high false alarm rate and only words with higher than 0.7 posterior probability would have a lower false alarm rate. For this system, using the posterior probability directly to construct the histogram may not be appropriate since for this system, we may wish to suppress all the words with posterior probability in the range of 0 to 0.7. In a counter example, if the recognizer is tuned for decoding speeds close to real-time the output lattice might be highly pruned. A shallow lattice has few competing hypotheses and therefore the arc posteriors might be over-estimated.

These examples show that arc posterior probabilities are system dependent and may not correlate well with the probability of the word being a true hit. However, one can try to learn a mapping from the posterior probability to the probability of correct detection. A non-parametric approach is to estimate the probability of correct detection, $P(\text{corr})$, according to:

$$P(\text{corr}) = 1 - \frac{\#\text{FA}}{\#\text{hits}} \tag{2}$$

Hence, given a development set, we can compute $P(\text{corr})$ for each bin of posterior probability where a bin represents a range of posterior probability. The bins can be of equal size or have the sizes depend on the number of words in the bins. For each bin, we can count the false alarms and the number of words

in the lattices of the development set and hence, we create a mapping from posterior probability to $P(\text{corr})$. Instead of using the posterior probability to construct the histogram, we can use $P(\text{corr})$ to construct the histogram. Alternatively, we can use a parametric approach to estimate the map or simply fit a parametric model to the non-parametric estimate of $P(\text{corr})$.

5. System Description

We tested our approach on a large benchmark dataset (MED11) used in the 2011 TRECVID MED Evaluation. Some events of interest (e.g. "Parade", "Parkour") do not contain any speech at all. Since the focus of this paper is to leverage speech content in videos, we restricted the list of videos and corresponding events to videos that contain at least some minimal amount of speech. The selected target events are: "Birthday party", "Changing a vehicle tire", "Getting a vehicle unstuck", "Making a sandwich", "Repairing an appliance" and "Sewing project". The selection was obtain automatically via a Speech Activity Detector (SAD) described below. The training set consists of 4852 videos containing 1200 videos from the 6 events of interest, with approximately 200 examples per event, and the rest of the videos are from the background class. Similarly, the test set consists of 2218 videos containing 410 videos from the 6 events of interest, with 65 to 75 examples per event.

We first transformed the raw audio into a 45 dimensional feature stream using the following steps. 14 Mel-warped cepstral coefficients were extracted from overlapping frames of audio data, each 29 ms long, at a rate of 100 frames per second. Each segment of speech is normalized by the mean cepstrum and peak energy non-causally, removing any long term bias due to the channel. In addition, the feature vectors were scaled and translated such that for each video, the data has zero mean and unit variance. These 14 base cepstral features and the energy, together with their first and second derivatives, compose the final 45-dimensional feature vector.

Then, within the video clips, the speech segments were identified by a SAD system. The SAD system employed two Gaussian mixture models (GMM), for speech and non-speech observations respectively. A small subset of 100 video clips was annotated for speech segments, which were used for training the speech GMM. Besides the non-speech segments in this set, we also used 500 video clips with music content to enrich the non-speech model, in order to handle the heterogeneous audio data in MED11. SAD was evaluated on 40 video clips and obtained a False Alarm rate of 10.1% and Miss Detection of 5.8% according to the NIST *md-eval* metric.

Given the automatic discovery of speech segments, we then applied BBN's large-vocabulary ASR system to the speech data to produce a transcript of the spoken content. This system was adapted from an ASR system trained on 1700-hour Broadcast News. In particular, we updated the lexicon and language model using MED11 descriptor files, relative web text data, and the small set of 100 video clips with annotated speech transcription. We've used a trigram language model trained over 2 million words of in-domain data from the MED11 descriptor files and relative web text data and 11 billion words of out-of-domain web-data. The acoustic models are adapted during ASR decoding for each video clip in an unsupervised fashion via Maximum Likelihood Linear Regression (MLLR) and Constrained MLLR (CMLLR). We evaluated the baseline ASR model and adapted ASR model on a held-out set of 100 video clips from the MED11 set. The WER of the baseline system was 48.2% and the WER of the adapted system was 35.8%.

Feature	Keyword List	AUC
1-best (0/1)	0/1	0.607
1-best (0/1)	*wc*	0.625
1-best (*wc*)	*wc*	0.670
cnet (*ap*)	*wc*	0.753
cnet (*pc*)	*wc*	0.755
cnet (*pc*) + word-clusters	*wc*	0.767

Table 1: Performance comparison in terms of AUC using different vector representations and keyword lists. Keywords are selected via the standard TF-IDF (0/1) or with word confidences (*wc*). Features are based on the 1-best with or without word confidences and confusion network (cnet) with arc posteriors (*ap*) or the $P(\text{corr})$ estimate (*pc*). The fusion with word-cluster features is also reported.

6. Experimental Results

In this section we are presenting the experimental results of our proposed methods. The event detection performance is measured in terms of the area under curve (AUC) metric. Table 1 summarizes the experimental results that will be discussed next.

6.1. Robust Data-Driven Keyword Selection

We first employ the robust data-driven keyword selection procedure proposed in Section 2 to derive the keywords. To do so, we apply the TF-IDF formula of Equation 1 and select the top-N keywords from each event. Table 2 illustrates examples of the automatically derived keywords using the approach proposed in Section 2. We can see that the selected keywords tend to be highly informative of the respective events.

We then compare the event detection performance under three configurations: (a) selecting and extracting features without using the word confidences (second row of Table 1), (b) selecting keywords by using the word confidences and extracting features without using the word confidences (third row of Table 1) and (c) using word confidences for both keyword selection and extraction (fourth row of Table 1). The results indicate that the use of word confidences for both keyword selection and extraction yields a 10.4% relative improvement. For the rest of the experiments we adopt the keyword list obtained with word confidences.

6.2. Confusion Networks

We then assess the benefit of looking beyond the 1-best hypothesis by leveraging the entire ASR output lattice. We extract the keyword hits and their associated posterior probabilities from the confusion network and constructed the histogram. Next, to estimate the probability of a keyword hit being correct, $P(\text{corr})$, we first use the small set of 100 video clips with annotated speech transcriptions to evaluate the keyword spotting performance using the keyword list from Section 6.1. Then, we sort all the keyword hits according to their posterior probabilities. We group them into different bins and ensure that each bin had at least 200 but no more than 1000 occurrences. Finally, we apply equation 2 to map all posterior probabilities to $P(\text{corr})$ and construct the histogram features. Figure 1 shows the mapping learned from the 100 video clips. We observe that the map suppresses keywords with high arc posterior probabilities and retains the scores of those with low posterior probabilities. This result indicates that keyword hits with arc posterior

Event	Keywords
Repairing an appliance	wire, washing, damaged, screw, screwdriver, dishwasher, circuit, thermostat, hinge, disconnect
Making a sandwich	sandwich, panini, bread, cheese, ham, bacon, pepperoni, mayonnaise, pesto, butter, lettuce
Changing a vehicle tire	lug, tire, wheel, tow, valve, hazard, wrench, jack, flat, change, puncture, spare, roadside
Sewing project	stitch, thread, trim, fabric, quilt, pleat, sew, zipper, pin, seam, dresden, scrapbook, design

Table 2: Examples of automatically derived keywords for the target events.

basil, cilantro, dill, herb, mint, mustard, oregano, thyme, tarragon
crab, crawfish, mussel, oyster, scallop, shellfish, shrimp, prawn
cayenne, coriander, cumin, ginger, paprika, sesame, turmeric
brunch, dine, dinner, lunch, luncheon, meal, supper
cheddar, feta, mozzarella, parmesan, ricotta
beef, chicken, lamb, meat, pork, poultry

Table 3: Examples of automatically derived word-classes related to the "making a sandwich" event derived via affinity propagation clustering using pair-wise distances from a thesaurus.

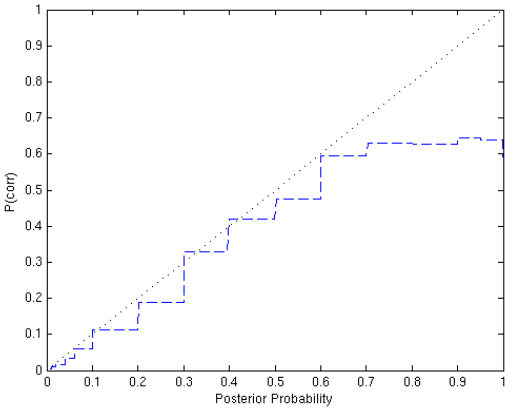


Figure 1: The mapping from posterior probability to $P(\text{corr})$ learned from the small set of 100 video clips.

scores greater that 0.7 are not necessarily more accurate. The mapping from the posterior probabilities to $P(\text{corr})$ provides an accurate representation of the relationship between the score for a particular word and its actual probability of being correct.

The experimental results show a 12.4% relative improvement by using keyword hits from the confusion network ("cnet (*ap*)" row of Table 1) compared to using the 1-best only. However, we obtain almost no improvement when we map the posterior probabilities to $P(\text{corr})$ ("cnet (*pc*)" row of Table 1).

6.3. Word Clusters for Keyword Expansion

We then derive word-clusters for keyword expansion according to the proposed method of Section 3. Using the proximity-based word pairwise similarities from the thesaurus [11] we apply the affinity propagation clustering algorithm to automatically generate the word clusters. Table 3 illustrates examples of classes related to the "making a sandwich" event. We observe that the clustering algorithm provides word clusters that are semantically related. Finally, by fusing the confusion network features (cnet(*pc*)) with the word-class features, via Multiple Kernel Learning (MKL) [12], we obtain an additional 1.6% relative improvement (last row of Table 1).

7. Conclusions

In this paper, we propose a comprehensive solution to the event detection problem in consumer-generated multimedia data by leveraging available spoken content. We present scalable, robust and data-driven methods that overcome several challenges found in web videos such as the absence of labeled speech audio content, noisy audio conditions, and sparsity of data. These proposed integrated approach improves, 26% relative, beyond the ASR-based features used in BBN VISER system for the 2011 TRECVID MED evaluation.

8. References

[1] P. Natarajan *et al.*, "BBN VISER TRECVID 2011 multimedia event detection system," in *Proceedings of NIST TrecVid 2011 Workshop*, Gaithersburg, MD., 12 2011.

[2] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Aug. 1988.

[3] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, pp. 455–470, 2005.

[4] S. Xie and Y. Liu, "Using confusion networks for speech summarization," in *HLT-NAACL*, 2010, pp. 46–54.

[5] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[6] L. Hollink, V. Malaisé, and G. Schreiber, "Thesaurus enrichment for query expansion in audiovisual archives," *Multimedia Tools Appl.*, vol. 49, no. 1, pp. 235–257, Aug. 2010.

[7] D. Lin, "Automatic retrieval and clustering of similar words," in *COLING-ACL*, 1998, pp. 768–774.

[8] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.

[9] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, pp. 373–400, 2000.

[10] G. Evermann and P. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *ICASSP*, vol. 3, 2000, pp. 1655–1658.

[11] http://webdocs.cs.ualberta.ca/~lindek/downloads.htm.

[12] S. Vishwanathan, Z. Sun, N. Ampornpunt, and M. Varma, "Multiple kernel learning and the SMO algorithm," in *Advances in Neural Information Processing Systems 23*, 2010, pp. 2361–2369.