AUDIO-BASED AFFECT DETECTION IN WEB VIDEOS

Dave Chisholm[1], Behjat Siddiquie[1], Ajay Divakaran[1], and Elizabeth Shriberg[2]

SRI International, Princeton, NJ[1] & Menlo Park, CA[2], USA

ABSTRACT

We present a new technique for detecting audio concepts in web content as well outline the technique's applications to video sequence parsing. Our focus is primarily on affective concepts and in order to study them we have collected a new dataset, consisting of videos where a speaker is persuading a crowd, called "Rallying a Crowd". We develop new classifiers for graded levels of arousal in speech as well as crowd noise and music and demonstrate their effectiveness on web content. These techniques achieve high detection accuracy (58.2%) for affective concepts on this new dataset and outperform (36.8%) state-of-the-art techniques (33.1%) for semantic concepts on a previously collected dataset. We also develop a new audio sequence segmentation technique which enables us to rapidly classify subsections of test sequence audio into the aforementioned audio classes. We are thus able to robustly address the detection of affective concepts in highly variable web content as well as the computational challenge of quick classification so as to enable web scale processing.

Index Terms— Audio concept detection, Audio segmentation, Affect detection

1. INTRODUCTION

Social multimedia has a powerful and swift impact because it enables wide and instantaneous dissemination of rich multimodal content. Computational extraction of attributes, scenes, objects and concepts from this content has been carried out with increasing success [1, 2], but in addition to this, social multimedia also often has strong affective content. The manner and affect of communication - for instance, calm speech versus agitated cries - strongly influences both the viewer reaction as well as any semantic interpretation of the multimedia content or its creator's intent. Therefore, affect detection is a key part of any attempt to automatically correlate multimedia content to its potential viewer response or to its influence across social networks. Our scope in this paper is to develop a content model for affect in the audio portions of social multimedia.

Extraction of affect from non-curated web content is still rudimentary because the bulk of automated affect extraction has focused on audio-visual content that has been captured in controlled conditions [3, 4, 5]. However, models trained on such carefully curated content may not perform well on web content. Therefore any content model suitable for goals such as estimating viewer response or predicting dissemination across social networks requires robust affect extraction that gives consistent results across the highly variable capture conditions inherent in web content.

In this paper, we propose a solution to an important part of the affect extraction problem viz. robust extraction of affect from the audio signal of web audiovisual content. We focus on a sub-genre of persuasive content consisting of a speaker rallying a crowd (RAC). This genre is rich in range and intensity of audio affect, and so is suitable to develop a bounded problem and testbed for this research. We develop for the first time classifiers for several grades of speech arousal that are robust across variable capture conditions and audio quality. We likewise develop robust detectors for related categories such as crowd reaction and music. These classifiers enable temporal localization of events of interest within such RAC videos such as a highly animated speaker or the call response pattern commonly observed between leaders and crowds during events such as rallies and protests. We also validate our approach to classification by applying our techniques to a pre-existing dataset, the *Columbia Consumer Video* (CCV) dataset [6] and comparing our results to the state-of-the-art (SOTA).

2. RELATED WORK

Concept detection in audio has long been an important problem in the audio processing community due to its applications in video retrieval. A standard approach to concept detection typically involves extracting audio features from the video, pooling them and using a classifier to recognize the different concepts [7, 8, 9]. Implementations differ in the details of strategies used for feature extraction, pooling and classification. Our concept detection takes a similar approach; however, its novelty lies in using multiple heterogeneous features and effectively fusing them. While we focus primarily on affective concepts, we demonstrate that our approach is general enough to detect selected semantic concepts as well.

Emotion Detection in human speech is rapidly increasing in importance because of its applications in enabling realistic human computer interactions. A standard approach to solving this problem involves extracting frame-wise low-level descriptors (LLDs) from the audio signal and then using functional features to aggregate these features over the utterance level [3, 4]. However, most of the work in this area has been demonstrated on datasets collected in controlled environments with little noise. In contrast, we work with unedited

videos from YouTube and show that our concept detection approach can be effectively used to recognize the emotional content in human speech.

Audio Segmentation is often used for separating the speech and non-speech segments (e.g. [10]) within audio. There has also been some work on unsupervised segmentation of an audio stream into semantically homogeneous segments. Notable among them is [11], which segments the audio signal by modeling it as a gaussian process and uses the Bayesian Information Criterion (BIC) to detect change points. A single video can contain multiple concepts and therefore instead of assigning a single concept to the entire video, a superior strategy is to segment the video in an unsupervised manner and then assign separate concepts to each segment [12, 13]. An additional advantage of this approach is that it provides video information at a finer granularity. In our work, we propose a multi-scale segmentation algorithm that allows for segmentation and concept labeling at multiple granularities.

3. APPROACH

Figure 1 illustrates an overview of our approach. The approach consists of feature extraction, codebook construction, quantization of the features into a bag-of-words representation based on the codebook, followed by learning and inference of a Multiple Kernel Learning based classification model. We describe each of these steps in detail below.

3.1. Bag-of-Words Representation

Feature Extraction From the audio signal we extract the Mel-Frequency Cepstral Coefficients (MFCC), the Spectrogram and Prosodic features. MFCCs and Spectrogram based feature representations have been commonly used for audio concept detection [6, 14]. Prosodic features include loudness, pitch, and speaking rate and they have been shown to be very effective for detecting emotional content in speech [15]. Our experimental results show that these features contain complementary information and that combining them leads to improvements in the concept detection performance. All of these features are extracted using the OpenSMILE toolkit [16]. Additionally, we also extract *Acoustic Unit Descriptors* (AUDs) [9], which model distributions of short audio sequences and therefore capture local temporal variations within the signal. The AUDs produce a quantized representation over short patterns of audio sequences.

Feature Encoding With the MFCC, Spectrogram and Prosodic features, we adopt the bag-of-words representation. Given features extracted from the training videos, we use hierarchical K-means to build vocabularies of size 10000. We construct separate vocabularies for energy normalized and unnormalized features, as we find that normalized features capture pitch based characteristics while unnormalized features encode loudness based characteristics. Once we have constructed these vocabularies, we quantize the features extracted from the training and test videos to obtain a histogram representation corresponding to each feature type. AUDs are already quantized by the nature of this feature and so no further encoding is needed for them.

3.2. Inference and Learning

One can compute a non-linear kernel corresponding to each feature representation and use a kernel SVM for classifying the video into one of the specified semantic categories. However, in order to effectively fuse these features for concept detection, we instead compute a composite kernel which is a convex combination of the individual kernels for SVM based classification. The kernel combination parameters are learnt using Multiple Kernel Learning (MKL) [17]. MKL jointly learns the kernel combination parameters as well as the SVM classification within the same optimization framework and has shown to be very effective at combining multiple heterogeneous features in several domains [18, 19].

Learning MKL Parameters: In MKL, the composite kernel is a convex combination of the basis kernels:

$$\Phi(x_i, x_j) = \sum_{k=1}^{K} d_k \phi_k(x_i, x_j), \quad \sum_{k=1}^{K} d_k = 1, \quad d_k \geq 0 \quad (1)$$

where x_i are the data points, $\Phi_k(x_i, x_j)$ is the kth kernel and d_k are the weights given to each information kernel. The optimization equation is given by:

$$\begin{aligned} \min \quad & \sum_k \frac{1}{d_k} w_k w_k^T + C \sum_i \xi_i \\ \text{subject to} \quad & y_i \sum_k \phi_k(x_i) + y_i b \geq 1 - \xi_i \ \ \forall i \\ & \xi_i \geq 0 \ \ \forall i, \ \ d_k \geq 0 \, \forall k, \ \ \sum_k d_k = 1 \end{aligned} \quad (2)$$

where b is the bias, ξ_i is the slack afforded to each data point and C is the regularization parameter. The solution to the above MKL formulation is based on a gradient descent on the SVM objective function. An iterative method alternates between determining the SVM model parameters using a standard SVM solver and determining the kernel combination weights using a projected gradient descent method.

Inference: During the training phase, we do not require segmentation as we can simply use the segmentations labeled by human annotators. On a test video, our inference consists of two steps: dividing the sequence into multiple segments and then classifying each segment by assigning it a concept label.

Step 1 - Multi-Scale Segmentation: The above mentioned approach for concept detection works well when we are certain that the entire video should have the same semantic label. While this is true in the case of the CCV dataset, in the RAC dataset, where the videos are significantly longer, assigning a single concept to whole video would lead to incorrect results. Furthermore, temporally segmenting the video and assigning a separate concept label to each segment provides deeper insights into its characteristics. Hence we first

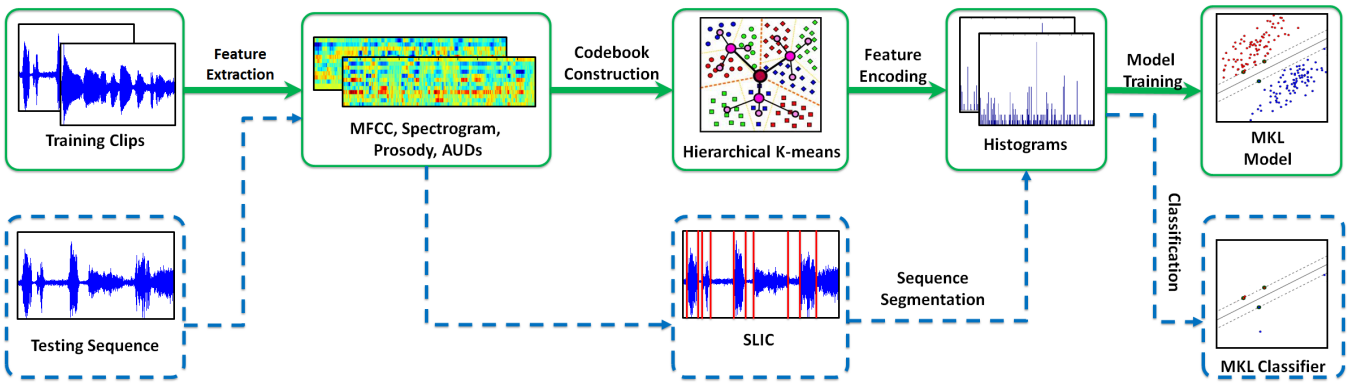

Fig. 1. Our approach for audio concept detection. The blue dotted line denotes the test pipeline, while the green solid line denotes the training process which also involves constructing the codebook and training the classifier.

segment the video into temporally homogeneous segments and then assign concepts to each segment. For segmenting the video into temporal segments based on its audio content, we adapt the Simple Linear Iterative Clustering (SLIC) proposed by Achanta et. al [20] for image segmentation to audio data. We initialize the cluster centers corresponding to each segment by sampling the audio frames at regular intervals. Next the initialized segments are iteratively refined in a two step process - the first step involves assigning each frame to a cluster-center in a local neighborhood by computing the distance in the feature space and the second step involves updating the cluster centers to reflect the new assignments. These steps are continued until the segmentation converges, and a post-processing step ensures temporal consistency by merging any non-contiguous cluster portions into a temporally adjacent cluster. SLIC has several advantages - it is unsupervised, extremely fast, and allows for segmentation to be done at multiple scales by varying the number and spacing of the initial cluster centers. It has been shown to work well for image segmentation and as we show in our results below, it also works well for audio segmentation using the spectrogram as the feature space. The segmentation allows us to bypass the need to efficiently search over all possible sliding windows by segmenting the audio into a set of homogeneous segments that are likely to contain a single concept.

Step2 - Classification: Given a new video, we extract all of the above features from each segment and quantize them in terms of their respective vocabularies. We compute the intersection kernel for each feature representation and use the weights learnt by MKL (Eq. 1) to compute a composite kernel. The composite kernel is then used for audio concept detection using the SVM parameters learnt in Eq. 2, thereby assigning a label to each segment.

4. DATASETS

Data Collection: A new dataset, called the *Rallying a Crowd* (RAC) dataset (Figure 2) was collected from YouTube for this research. The entire dataset consists of 132 positive and 59 negative example videos with an average duration of 7 minutes. Positive examples contain instances of a speaker vehemently trying to rally or persuade a crowd. They include material such as heated political debates as well as professional and amateur recordings of controversial events such as protests or rallies. Such videos tend to consist primarily of spoken content of varying arousal levels from a single speaker along with crowd response and background music. Negative examples consist of videos that are semantically similar but affectively different from positives (e.g. political interviews and lectures) or semantically different with some shared traits (e.g. a stand up comic addressing a crowd). A wide variety of languages were spoken in the videos. The RAC dataset's focus on affectively persuasive political content is unique in the field.

Fig. 2. Public domain images representative of RAC.

Data Annotation: For the experiments below, 102 positive videos were each annotated by the same pair of human subjects. Each annotator first divided an entire audio track into disjoint segments that were semantically homogeneous based on their audio content. Segments could be any length, but could not overlap and had to cover the entire audio track. The annotator then assigned each segment one of the 9 different audio categories listed in Table 1, or marked a segment as "Ignore" (e.g. silence, static, or otherwise not covered by the classes below). Annotators only listened to audio content, and did not actually view the video, to prevent visual cues from affecting their judgment of the audio content. On average, semantic segments were each about 15 seconds long.

Difficulties inherent in semantic and affective segmentation became apparent during the annotation process. The original set of concepts included 16 audio categories; this was

1. Crowd	6. Calm Speech
2. Music + Crowd	7. Slightly Agitated Speech
3. Music	8. Agitated Speech
4. Music + Speech	9. Very Agitated Speech
5. Crowd + Speech	

Table 1. Audio Categories in the Rallying a Crowd Dataset

empirically reduced to 9 by examining confusion matrices between the annotators. Notably, different types of crowd noise were often confused (e.g. cheering vs. clapping). Across all annotated audio (645 minutes) both annotators assigned the same class only 54.5% of the time (351 minutes), implying significant disagreements on classification even among humans. While some of the disagreement can be ascribed to use of labels on a continuum (e.g. "slightly agitated" vs. "agitated") or disagreement over the exact start and end time of segments, a ground truth for this type of content classification is difficult to define. To reduce the impact of this ambiguity on our experiments, we used only portions of segments where both the annotators agreed on the concept label.

Prexisting Datasets: Experiments were also run against the CCV dataset [6]. It consists of 9317 unedited videos from YouTube, each of which is been labeled with one of 20 semantic categories. These categories are very diverse and range from objects (e.g., "cat" and "dog"), scenes (e.g., "beach" and "playground") to events (e.g., "baseball", "skiing", "graduation" and "music performance").

5. EXPERIMENTS

5.1. Concept Detection on CCV Dataset

We first performed concept detection on the CCV dataset and compared to SOTA results to validate our approach. We follow the experimental protocol in [6] using the same train and test splits. We build the codebooks for each feature type, learn the classifiers on the training data, and report the Mean Average Precision (MAP) scores on the test data. Note that here a single concept label is to be assigned to the entire video and hence no segmentation is required. Our aim is to show that our MKL based concept detection approach is general enough to detect both semantic and affective concepts.

Table 2 shows the MAP scores obtained from different feature combinations using MKL. Even in case of the individual features, we combine the kernels induced by the normalized and unnormalized features as well as the AUDs, using MKL. The results demonstrate that the features contain complementary information and combining them leads to a significant increase in performance.

Table 3 shows that our approach significantly outperforms previously published audio-only results on this dataset, demonstrating the advantages of using multiple feature types over a single feature [6, 14] [1].

[1]While [6, 14] use the same overall approach, we believe that the superior results of [14] are due to their better implementation and parameters.

Features	**MAP**
Prosody	21.7
MFCC	26.9
Spectrogram	27.8
AUDs	32.2
All Features	**36.8**

Table 2. Performance of different features as well as fusion of all of the features using MKL on the CCV dataset.

Approach	**MAP**
MFCC + non-linear SVM [6]	28.3
MFCC + non-linear SVM [14]	33.1
Multiple features + MKL (proposed)	**36.8**

Table 3. Comparison with previously published results on the CCV dataset showing our improvement over the SOTA.

5.2. Concept Detection on RAC Dataset

We performed similar experiments on the RAC dataset by dividing the dataset into "snippets" of a known class. The creation of snippets segmented the audio tracks into semantically homogenous sections and also dealt with annotator disagreement by excluding sections where there was not a unanimously chosen label. Snippets correspond to regions where both annotators labeled temporally overlapping segments as the same class; the overlapping portion was extracted as a snippet and tagged with the unanimous label. 737 snippets were extracted from training videos and used to create the concept detectors, while 671 snippets were extracted from test videos and used for evaluation. Table 4 shows the results. Again combining multiple features results in a substantial performance increase, demonstrating that MKL is very effective at fusing information from multiple hetereogeneous features.

Features	**MAP**
Prosody	48.8
MFCC	51.1
Spectrogram	51.3
AUDs	42.7
All Features (MKL)	**58.2**

Table 4. Performance of classification using different features (and all features fused using MKL) on the RAC dataset.

The results in Tables 2,3,4 demonstrate that our approach is general enough to both effectively detect affective concepts on the RAC dataset, and also outperform the current SOTA in detecting semantic concepts on the CCV dataset.

5.3. Quantitative Results on Automatic Segmentation and Concept Detection

We now evaluate the performance of segmentation followed by concept detection on the RAC dataset. As described in subsection 3.2, we use SLIC to segment each video at three

different empirically chosen scales (fine, medium, and coarse) with average segment length of 1.3 seconds, 4.1 seconds and 15 seconds respectively. Each segment is then independently classified based on the individual features as well as their combinations using MKL. Due to disagreements between annotators, there is a lack of an authoritative ground truth for some portions of the videos; it is ambiguous how to evaluate the accuracy of a machine labeling for portions of audio where even the two human annotators disagreed. As a metric we evaluate only the portions of video unanimously annotated, and provide as "accuracy" the percentage of such material where the machine label matched the unanimous human label. We ignore portions where the human annotators disagree on the labels; this is consistent with how the training data was generated and also effectively removes some of the more difficult cases. Table 5 shows the results. Here again we can observe that while the performance of individual features varies, fusing them using MKL results in a consistent and significant improvement in performance. While the overall performance at different scales is similar, there are differences in the detection rates of the individual classes at different scales. This is discussed further in subsection 5.4.

Features	**Acc. Fine Scale**	**Acc. Medium Scale**	**Acc. Coarse Scale**
Prosody	45.3	45.1	44.9
MFCC	43.4	44.3	44.0
Spectrogram	41.4	42.7	41.4
AUDs	38.7	38.3	35.4
All Features	**47.3**	**48.9**	**47.4**

Table 5. Classification accuracy (Acc.) of different feature combinations for segmentation followed by concept detection at multiple scales on the RAC dataset.

In order to further evaluate the effectiveness of our SLIC based segmentation we compared its performance against a uniform segmentation at the same scale. Each video was uniformly segmented into fixed width segments of 1.3, 4.1 and 15 seconds so as to match the segmentation scales of SLIC. The results (Table 6) demonstrate that SLIC outperforms a uniform segmentation at all scales with larger improvements at finer scales. While the improvements of around 2-3% seem small, we note that the snippet classification accuracy of 58.2% (Table 4) represents performance given an "ideal" segmentation (i.e. a segmentation where any duration in which both human annotators agreed on labeling is grouped into a single temporally contiguous segment). Thus we believe the improvement due to SLIC is significant, since it represents ˜20% of the difference between classification given a uniform segmentation versus an ideal segmentation.

5.4. Qualitative Results on Automatic Segmentation and Concept Detection

Figure 3 displays human annotation as well as automatic results on footage of Martin Luther King's *I Have a Dream* speech. While Table 6 seems to indicate very similar classification performance regardless of SLIC segmentation scale, in fact there were apparent performance differences dependent on content type. For instance, the finer scales resulted in improved performance classifying crowd noise and better captures the "call and response" characteristic common in these videos – long periods of oration punctuated throughout with brief and vehement crowd response. This phenomenon is clear from time 4:00 to 14:00 in this video. The coarse scale resulted in better performance classifying music as can be seen at the start of this example, during which the crowd is singing the protest song, *We Shall Overcome*. These characteristics of "call and response" and a musical introduction were frequently observed in RAC videos.

Segmentation	**Acc. Fine Scale**	**Acc. Medium Scale**	**Acc. Coarse Scale**
SLIC	**47.3**	**48.9**	**47.4**
Uniform	43.9	46.8	46.8

Table 6. Classification accuracy (Acc.) for SLIC vs. uniform segmentation (using all features in both).

6. CONCLUSION

We have proposed an approach for affective and semantic concept detection in web content along with an audio segmentation approach to enable video sequence parsing. The proposed approach outperforms the SOTA on semantic concept detection on the CCV dataset and produces good results for video segmentation and affective concept detection on the newly collected RAC dataset.

In the future we intend to improve the RAC dataset by enlarging it and investigating the issues of annotator disagreement. We also wish to extending our approach to include other modalities besides audio and work towards automatically detecting persuasive content in web videos.[2]

7. REFERENCES

[1] A. Tamrakar et al., "Evaluation of low-level features and their combinations for complex event detection in open source videos," *CVPR*, 2012.

[2] P. Natarajan et al., "Multimodal feature fusion for robust event detection in web videos," *CVPR*, 2012.

[3] D. Littman and K. Forbes, "Recognizing emotions from student speech in tutoring dialogues," *ASRU*, 2003.

[4] B. Schuller et al., "Avec 2011 -the first international audio visual emotion challenge," *ACII*, 2011.

[2]This material is based upon work sponsored by the Defense Advanced Projects Agency under the U.S. Army Research Office Contract Number W911NF-12-C-0028, through IBM Corporation subcontract 4914004308. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the position or policy of the U.S. Army Research Office, DARPA, DoD and IBM Corporation and no official endorsement should be inferred.

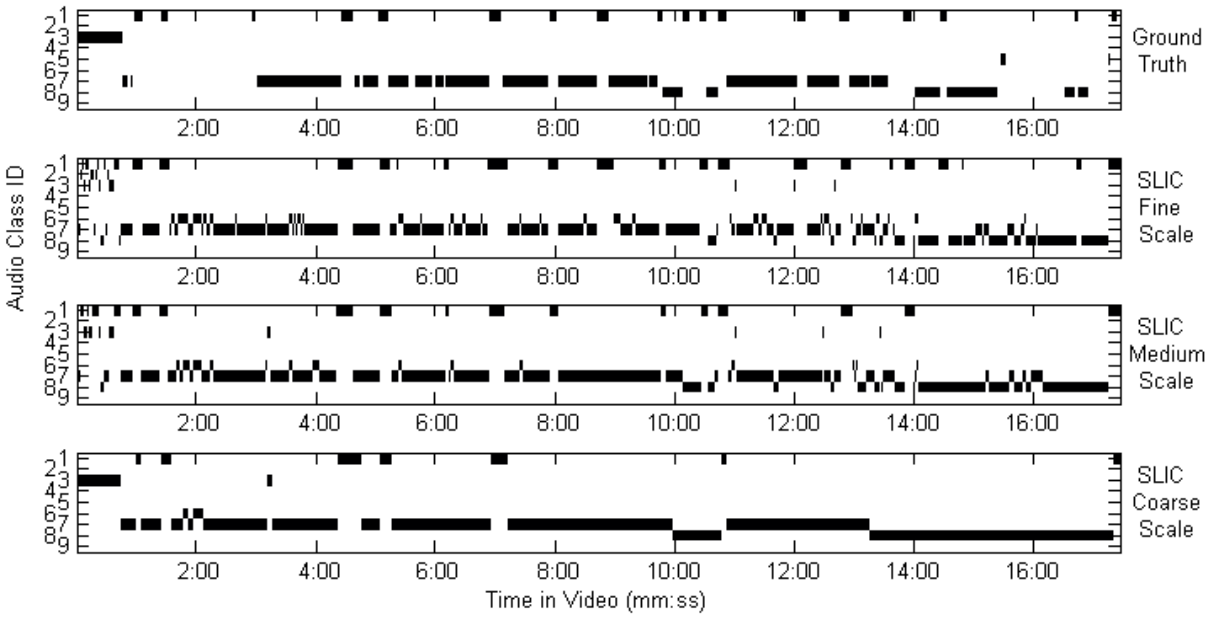

Fig. 3. Human and automatic segmentation and classification for Martin Luther King's *I Have a Dream* speech. Top plot contains segments where two human annotators agreed on content classification; time where no class is indicated corresponds to periods where annotators disagreed on content class. Lower three plots contains segmentations using increasing large scale parameters in the SLIC algorithm (average segment sizes 1.3, 4.1 and 15 seconds respectively) and the resulting classifications. Audio class ID 1 is Crowd, 3 is Music and 6-9 are speech with increasing levels of agitation (See Table 1 for full list). Video may be viewed at https://www.youtube.com/watch?v=smEqnnklfYs

[5] M. Amer et al., "Emotion detection in speech using deep networks," *ICASSP*, 2013.

[6] Y. Jiang et al., "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," *ICMR*, 2011.

[7] S. Chang et al., "Large-scale multimodal semantic concept detection for consumer video," *ACM MIR*, 2007.

[8] K. Lee and D. Ellis, "Audio-based semantic concept classification for consumer video," *IEEE TASLP*, 2010.

[9] S. Chaudhuri et al., "Unsupervised learning of acoustic unit descriptors for audio content representation and classification," in *INTERSPEECH*, 2011.

[10] M. Graciarena et al., "All for one: Feature combination for highly channel-degraded speech activity detection," *INTERSPEECH*, 2013.

[11] S. Chen et al., "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 127–132.

[12] D. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *ACM Workshop on CARPE*, 2004.

[13] T. Zhang et al., "Audio content analysis for online audiovisual data segmentation and classification," *Speech and Audio Processing, IEEE Transactions on*, 2001.

[14] Y. Jiang, "Super: Towards real-time event recognition in internet videos," *ICMR*, 2012.

[15] J. Ang et al., "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *ICSLP*, 2002.

[16] F. Eyben et al., "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *ACM MM*, 2010.

[17] F. Bach et al., "Multiple kernel learning, conic duality, and the smo algorithm," *ICML*, 2004.

[18] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," in *JMLR*, 2008.

[19] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," *ICCV*, 2009.

[20] R. Achanta et al., "Slic superpixels compared to state-of-the-art superpixel methods," in *IEEE PAMI*, 2012.