There is No Data Like Less Data: Percepts for Video Concept Detection on Consumer-Produced Media

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

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

Howard Lei
International Computer Science Institute
1947 Center Street
Berkeley, CA 94704, USA
hlei@icsi.berkeley.edu

Ajay Divakaran
Stanford Research Institute
201 Washington Road
Princeton, NJ 08540, USA
ajay.divakaran@sri.com

ABSTRACT

Video concept detection aims to find videos that show a certain event described as a high-level concept, e.g. “wedding ceremony" or “changing a tire". This paper presents a theoretical framework and experimental evidence suggesting that video concept detection on consumer-produced videos can be performed by what we call “percepts", which is a set of observable units with Zipfian distribution. We present an unsupervised approach to extract percepts from audio tracks, which we then use to perform experiments to provide evidence for the validity of the proposed theoretical framework using the TRECVID MED 2011 dataset. The approach suggest selecting the most relevant percepts for each concept automatically, thereby actually filtering, selecting and reducing the amount of training data needed. It is show that our framework provides a highly usable foundation for doing video retrieval on consumer-produced content and is applicable for acoustic, visual, as well as multimodal content analysis.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Retrieval**]: Content Analysis and Indexing; D.2.8 [**Multimedia Activity and Event Understanding**]: Metrics:complexity measures, performance measures

General Terms

Theory

Keywords

Video Concept Detection, Theory, Audio

AMVA'12, November 2, 2012, Nara, Japan.

1. INTRODUCTION

Social network applications have transformed the Web into an interactive sharing platform where users upload multimedia documents, comment on, and share this content with the public. This, together with the advent of handheld recording systems carried by people almost all the time capturing video of anywhere with the press of a button makes the amount of consumer-produced multimedia documents increase rapidly on a minute-to-minute basis. However, all of these documents are of little value if users cannot retrieve them easily. Therefore, higher-level search paradigms, like video concept detection, are in increasing demand by a variety of users.

Video concept detection aims to find videos that show a certain event described as a high-level concept, e.g. “wedding ceremony" or “changing a tire". In contrast to broadcast TV, movies, songs, and other professionally-produced content, consumer-produced multimedia documents have a high variance in quality and content and typically do not obey any particular content format. Therefore, traditional retrieval approaches that rely on manual definition of predefined object detectors such as “face” or “walking person” mostly fail as these manual event definitions are very domain specific.

In contrast to surveillance videos, consumer-produced videos usually contain video as well as audio, making it possible to treat the video retrieval problem as a multimedia problem. Moreover, while retrieval problems in the past often suffered the problem of limited training data – thus the common saying “there is no data like more data" for training – consumer-produced videos are available in abundance, often paired with meta-data (such as geo-location or tags) that can be used as ground truth. This allows for machine learning algorithms to be trained with as much data as there is CPU time and memory.

The problem, however, is that with consumer-produced media having little to no structure, most of the data is noise, so that burning more CPU cycles might not help at all. In other words, we are entering an era where the paradigm “there is no data like more data" is changed to “there is no data like the right selection of data to train on" or in the end, as we suggest in the title, “there is no data like less data". So the major question becomes how to select the right data, i.e. how to reduce the abundance of data to a tractable amount of data that is maximally useful to build models for search. Decide how to define an event by its most representative acoustics.

Using the example of video concept detection, this paper investi-

gates a theoretical framework for the selection of the right training data from consumer-produced videos and presents experimental results that provide evidence for the validity of the framework based on large dataset of consumer produced videos. Our framework uses the notion of "percepts" which represent perceptually similar units. We describe a set of rules for the properties of the percepts and their mapping to concepts. Furthermore, based on both, empirical evidence as well as without loss of generality, we assumed the precepts follow a Zipfian distribution. This allows us to define the upper and lower bounds for the descriptiveness of the extracted percepts. We then derive some real-world conclusions backed up by experience in the community before we describe our own experimental evidence based on the NIST TRECVID 2011 MED corpus.

The paper is structured as follows. Section 2 starts presenting some related work. Section 3 then presents the theoretical framework before and overview of the system in Section 4 and then drawing some real-world conclusions in Section 5. Section 6 presents the data set and experimental setup before Section 7 describes an outline of the implementation of our system. Section 8 then continues with an analysis of the results in the scope of the framework. Finally, Section 9 resumes with an outlook into the future.

2. RELATED WORK

Our theoretical approach generalizes from practical work already published in the multimedia community. A good overview is presented at by [6] and [11]. Many approaches employ supervised learning techniques in which classifiers are trained to discover distinct low-level concepts such as "indoor/outdoor" or "people laughing". The number of classifiers trained in these approaches ranges from ten as described in [5] to 75 in an approach for video scene segmentation described in [10].

The downside of these supervised approaches is that training data has to be manually selected for each new low-level sound-concept in order to train models for new application domains and that each low-level sound category has to be anticipated by the specialists that train the system. We will comment on that further in Section 5. This was also extensively discussed in [3]. Only very few approaches take a more holistic perspective and use the entire audio contents of each file [9] and [2]. These papers can be interpreted as the first set of evidence that less data is actually more data when it comes to consumer-produced content, as their main limitation was the training on noise.

Approaches similar to our system implementation are described in [7] and in [2] although this article provides more theoretical grounding. The system in [7], however, is already inspired by term frequency and inverse document frequency.

3. CONCEPTUAL FRAMEWORK

3.1 Philosophy and Terms

We start by the definition of a percept, quoted from Merriam Webster's dictionary.

Definition 1. Percepts: an impression of an object obtained by use of the senses.

Since we are dealing with video concept detection, we will have to restrict ourselves to events. An event is a complex activity occurring at a specific place and time, involves people interacting with other people and/or objects, consists of a number of human actions, processes, and activities that are loosely or tightly organized and that have significant temporal and semantic relationships to the overarching activity, finally it is observable by human senses. Video concept detection is defined as the task of finding videos that describe the same concept as a set of example videos. Concepts are usually defined as a higher-level events, e.g. "building a shelter" rather than individual objects. In the end, there seems to be no crisp definition of a concept in the research community. Therefore, we decided to adopt one. We found the definition by philosopher John Locke (1632-1704) quite handy:

Definition 2. "A concept is created by abstracting, drawing away, or removing the uncommon characteristics from several particular individual ideas or observations. The remaining common characteristic is that which is similar to all of the different individuals."

The above definition actually tells us how we get from an observed event to a concept. However, philosophy even tells us more. Immanuel Kant (1724-1804) is famously quoted with:

OBSERVATION 1. *Concepts without percepts are empty; percepts without concepts are blind*

In other words: Observations without a mapping to a concept are considered noise and every concept needs to have at least an observation or it is empty. This gives us a notion of how to distinguish representative percepts from not so representative percepts. Also, it means we can only work on concepts that are actually representable.

3.2 Formalization

Let P be a set of percepts $\{p_1, p_2, ..., p_n\}$. Let C be a set of concepts $\{c_1, c_2, ..., c_n\}$. According to the definitions above, we then have a relation between the percepts and the concepts that we call language $L \subseteq P \times C$. We observe the following properties of L.

- A unique tuple $(p_i, c_j) \in L$ where only one exact p_i maps to only one exact c_j is called perfect or clear mapping.
- Tuples $(p_i, c_j) \in L$ where several p_i occur together with the same particular c_j are called synonyms.
- Tuples $(p_i, c_j) \in L$ where one particular p_i occurs together with several c_j are called ambiguous or homonymous.

Percepts can be defined as empty by adding a nil concept to C. Practical evidence suggests that we should also mention the case where several percepts always occur together to describe the same concept. We call this as a paraphrastic relation $PR \subseteq L$.

Definition 3. For a concrete $c \in C$, a model $M_c \subseteq L$ is defined as a function $M : P \rightarrow c$.

We will call a model complete when all percepts ever observing a concept are included in it. Since given a set of percepts, a model M can map to a concept, practically, our goal is to find a model for each concept and percept. Ideally, we want the model to have certain properties.

3.3 Properties of Models

As of Definition 2, the perfect model is one that contains all percepts common to one concept and only those. This means, we can define the purity of two models as a purity function. For example, the following function: $purity(M_1, M_2) = \frac{|p \in M_1|}{1+|p \in M_1 and M_2|}$ for all $p \in M_1$ is a purity function where higher values mean higher purity. Of course, other definitions are possible and for practical purposes actually many distance measures are possible, such as *KL* divergence or the inverse document frequency as used in [7] and other works. Therefore, we will not strictly adhere to the definition, which only serves as an example as of how purity could be defined most simply.

On a given set of example videos for one concept, a very important notion is the distribution of percepts occurring for one concept. The easiest way to estimate it is to obtain the frequency of occurrence of each percept. Based on empirical evidence reported in related work, see for example [9] or [3], we can assume that the distribution is Zipfian. Moreover, given the loose definition of percepts (see Definition 1), assuming a particular distribution can be done without loss of generality! It means in practice, that percept extraction methods should be performed with a certain target distribution in mind and in the following we will argue that this should be a Zipfian distribution.

Zipf's law originally states that given a corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Likewise, we assume that for a given concept, the frequency distribution of percepts is defined by the function:

Definition 4. $f(k, s, N) = \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)}$

with N being the number of percepts, k their rank (sorted from highest frequency to lowest), and s the value of the exponent characterizing the distribution which for our purposes can be defined as 1. In other words, the function defines that out of N percepts, the frequency of percepts of rank k is $f(k; s, N)$. While the value of f is not of practical matter given that, in practice, it's known and easily obtainable. However, the properties of Zipfian distributions are well understood.

For example, the cumulative distribution function of a Zipfian distribution is given as:

OBSERVATION 2. $CDF(k, s, N) = \frac{H_{k,s}}{HN,s}$

reusing the same variables as in the previous definition and $H_{n,m}$ being the n-th generalized Harmonic Number given by $H_{n,m} = \sum_{k=1}^{n} \frac{1}{k^m}$.

An immediate result of assuming a Zipfian distribution is that we can make a quantitative statement of how much of the concept is described by the first few n-ranked percepts under the assumption of a complete and pure model.

The lower frequent percepts are by this definition less descriptive. At the same time, the lower descriptiveness is paired with an equal chance of being ambiguous, making them better candidates for noise. With model purity being a goal for distinctiveness, we therefore deduct that it is better to train models with less data but at the same time higher descriptive power.

4. SYSTEM'S TECHNICAL OVERVIEW

We provide a description of the system which comprehends three main stages: Diarization, Clustering, and a TFIDF Filtering that feeds a Support Vector Machine.

4.1 Diarization

Speaker diarization systems were initially used to detect speakers and when they were speaking. In this case, we tuned a diarization system and used it to explore and detect the sounds that describe the video's content.

The first step for diarization is the Feature Extraction. The speech signal is parametrized in frames with a window size of 30 ms and a step size of 10 ms, computing 19 parameters, using Mel Frequency Cepstral Coefficients MFCCs.

For the second step, usually diarization systems apply a speech activity detection procedure to improve the detection of speech and discard the rest of the sounds in the audio. For our algorithm we use a speech/non-speech segmentation system that keeps the sounds, enhancing the performance of the segmentation and clustering step of the diarization.

For the last step is the core of the diarization, which is the segmentation and clustering with a "bottom-up" agglomerative hierarchical clustering, which means that we start with a large number of clusters that are gradually merged to improve some chosen metric, using some stopping criterion to determine when to discontinue merging. For this paper we used the Bayesian Information Criterion. The following outline describes the diarization algorithm:

1. An initial segmentation is generated by uniformly partitioning the audio in same length *S* segments. For speaker diarization the number of *S* is bigger than the assumed number of speakers, for example 16 for a normal meetings corpus. For this task we tuned it with 64 initial segments due to the diversity of sounds encountered in the TRECVIDMED 11 data. For each segment we trained a Gaussian Mixture Model GMM using the Expectation Maximization algorithm. We ensured a minimum duration of 200 ms to detect smaller duration sounds.

2. A re-segmentation is performed running a Viterbi decoder using the current set of GMMs to segment the audio.

3. The models of the current segmentation are re-trained.

4. Select the closest pair of clusters and merge them. This is an iterative method and in every iteration all the possible cluster combinations are checked computing the difference between the sum of the BIC scores of each model and the new model trained after merging a cluster pair. The clusters with the largest improvement in the BIC scores are merged and the new GMM is used. The algorithm repeats from step 2 until there are no remaining pairs that will improve the BIC scores.

4.2 Clustering

The output of the Diariazation are GMMs representing each of the representative "sounds" of the audio clip. The GMMs have three parameters, weight, mean and variance. All of them are combined to create simplified super vectors. The compilation of super vectors are fed to a K-means algorithm. The parameters chosen were previously tuned for this algorithm and are: a random seed selection based on the input data, 10 iterations and 300 clusters.

4.3 TFIDF Filtering and SVM

Once we have our output clusters we proceed to relate them to the diarization GMM super vectors of each audio. Each super vector corresponds to a sound identified by the diarization and it will be represented by the closest K-means centroid. The result of this is an abstract representation of the audio based on the 300 clusters of the K-means. The distribution frequency of the supervectores based on the 300 percepts resembles a Zipfian distribution which sometimes is related to the application of TFIDF techniques. The abstract representation for each file is a vector corresponding and representing each of the audio for each of the concepts. Each vector has the a dimension of 300 by 1, where we show the number of occurrences of the 300 percepts for each audio file.

The classification is done by the Support Vector Machine, in this system we used the LIBSVM implementation from [1]. We trained an SVM model with our compilation of the above-mentioned vectors. They don't always contain occurrences for all of the percepts. Therefore we create sparse vectors to feed our SVM. We also used an intersection kernel for a multi class concept classification using the concept labels E001-E005. A cross validation to tuned the learning options was performed too.

These vectors first go through a weighting step given by the TFIDF numerical statistics. The Term Frequency-Inverse Document Frequency show how important is a word to a document in a certain corpus. In order to understand the TFIDF approach we can do an analogy to text analysis, so each of the diarization output sounds correspond to a ‘’word" and each of the audio files correspond to a document of a corpus. Term Frequency is the number of times a word appears in that document, and can be defined by the equation 1.

$$TF(c_i, D_k) = \frac{\Sigma_j n_j(c_i = c_j | c_j \in D_k)}{\Sigma_j} \quad (1)$$

Where $TF(c_i, D_k)$ is the Term Frequency of audio word c_i in the audio document D_k. The other term is Inverse Document Frequency and tells you whether a word is common or rare across the documents and can be defined by the equation 2.

$$IDF(c_i) = log \frac{|D|}{\Sigma_k P(c_i \in D_k)} \quad (2)$$

Where |D| means the total number of documents and $P(c_i \in D_k)$ is the probability of term c_i in the document D_k.

Each concept or event is described by a different combination of these 300 clusters resulting from the total of videos representing that concept. We also analyzed the sounds related to the Top-TF most frequent K-values from the training set, and we found out that they were closely related to variations of speech and music, which means that these are the top frequent sounds among the employed training data set.

The entire training and testing audio datasets went through these above-mentioned steps in parallel. Except the testing data which wasn't fed to the SVM to trained the model, instead it was fed to the SVM to be classified along with the SVM model. The SVM provided us with a score for each of the 5 concepts, and the highest value was the assigned event label.

In this paper we experiment with the top-n percepts taking advantage of the understanding of how they describe each of our concepts.

5. REAL-WORLD IMPLICATIONS

The above-defined framework has several practical implications. Unfortunately, the given page limit for this article allows us only to present a few.

One conclusion of our framework is that the goal should be to find complete and pure models. The concrete definition of a percepts and concept don't actually matter. Given a Zipfian distribution of the percepts, however, a complete model can be approximated by finding maximally pure models of the top-n frequent percepts.

Since the definition of the percepts does not actually matter, the language can also be arbitrary. Of course, it is possible to switch out our definition of precepts by any other definition, eg. to use percepts that are more human-like. For example, if the percepts are defined to be words of English language then the problem of video concept detection can be mapped to natural language understanding. However, practically, it is hard to actually have the computer extract human-like percepts.

The standard approach for estimating human percepts is discussed in the related work: Classifiers that detect pre-trained objects and events are stacked together to find particular predefined percepts. The limits of the approach, however, are reached quickly when dealing with “wild" videos. This is also extensively discussed in [3].

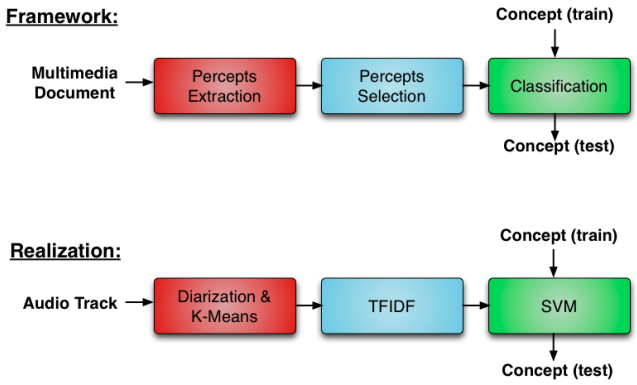

Figure 1: *Overview of the video concept detection system used for the experimental setup*

However, from our framework we can actually infer that it is completely unnecessary to define a concrete language a-priori or to define the exact concepts other than using a set of example videos that approximates a complete model. The corresponding real-world observation would be first-language acquisition or immersive learning experiences for foreign language acquisition (e.g. as utilized commercial software like Rosetta Stone) where models are automatically learned by a human without the need to map back to another language. Most importantly, as we know, word-concept mappings never match perfectly between two languages, making translation difficult. More than just providing an idea on how video concept detection can work, Observation 2 gives an upper limit for the descriptive power of a subset of percepts. It's an upper limit because it is only an actual limit when the models are pure and complete. Section 8 will elaborate on an example.

6. EXPERIMENTAL SETUP

We use the NIST TRECVID Multimedia Event Detection dataset (MED) DEV-T subset from the 2011 evaluation task. The recordings consist on multimedia content uploaded by public users with an average duration of three minutes. The organization includes 15 classes with five of those in the test set. The training set comprehends 2040 videos, and the test set 4251 where 492 of them belong to the five classes and the rest correspond to a random video category. The five classes used are: E001 Board tricks, E002 Feeding an Animal, E003 Landing a Fish, E004 Wedding, E005 Woodworking and other. The “other" class consists of all videos that do not belong to the first five classes. In this paper, we only use audio track of the videos. However, our framework can potentially be combined with video or used on visual data only as, again, the notion of percepts is very general (see Definition 1).

7. SYSTEM IMPLEMENTATION

We realized our audio percepts extraction by generalizing from a speaker diarization system. The approach is described in detail in [8]. The time-component of the diarization system helps consolidating paraphrastic percepts.

Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from the video soundtrack. The diarization system uses a temporal agglomerative hierarchical clustering approach to generate audio segments of similar acoustic structure. The segments are then clustered and represented using Gaussian Mixture Models (GMMs). In order to match the audio clusters across the different training videos belonging to one concept, the system reduces the GMMs to a single vector that consists of the sums of the weighted means and the

sums of the weighted variances of each Gaussian (in accordance with speech-community terms, we call this vector a simplified super vector). A K-Means method is then used to cluster the simplified super vectors, resulting in clusters that represent prototype clusters, which we define as being our percepts.

The entire video concept detection system based on the framework outlined above is show in Figure 1 in comparison with the conceptual approach outlined in previous sections. The diarization and K-means step represents the percepts extraction. Each concept is represented by 300 percepts, which we used to perform the experiments described in the next section. The GMMs corresponding to the percepts are then used to detect the same percepts in the audio tracks of the test videos. This allows a direct mapping comparison between the percepts in the training and test set. The top-n percepts selection follows a TFIDF approach [7] and a Support Vector Machine is used to perform the final classification. The SVM classification is described in detail in [4], however, without the proceeding steps.

8. ANALYSIS OF RESULTS

Top-N	Actual Hits	Predicted Hits	$Error$	Ambiguity
1	17 %	16 %	1 %	0 %
3	35 %	30 %	5 %	0 %
5	46 %	36 %	10 %	20 %
10	56 %	46 %	10 %	24 %
20	84 %	57 %	27 %	27 %
40	99 %	68 %	31 %	31 %

Table 1: Predicted descriptiveness of the top-n percepts for different values of n in comparison to the measurements on TRECVID MED11 as well as observed model impurity in %. While it remains unclear how complete the models are, our theoretical prediction is roughly within the range of actual hits considering the ambiguity.

As a first experiment, we want to verify that our framework actually can estimate the descriptiveness for the top-n percepts. We focused our experiments on a smaller set of videos containing the same 5 classes for train and test, with a total of 662 clips of training data, and 492 clips of test data. As discussed in the previous Section, we extracted 300 percepts for each concept in the training set. After that, we extracted 300 percepts for each test video. In this experiment, we control for the class both in test and training set in order to be able to match the percepts perfectly between test and training.

We then wanted to know how many videos in the test set actually match the training percepts in their perspective class, given a reduction to the top-n highest frequent percepts. In other words, assuming the training videos in MED 11 as a complete set of percepts for each concept (perfect model), we show how well are the percepts represented in the test set. Table 1 shows the results of the experiment.

We show the predicted descriptiveness of the top-n percepts for different values of n as calculated using Observation 2 and compare it to the empirically observed data. The measurements are obtained by counting the matching videos for all concepts vs. the non-matching. In other words, the top-1 column shows how many videos could in theory already be classified just based on the top-1 occurring percept. The ambiguity is determined by counting the number of homonymous percepts. As can be seen, the prediction error is pretty low and correlates with the ambiguity, which provides evidence for the validity of our framework. Please note, real-world audio percepts only approximate the Zipfian distribution and the models are not complete.

As a second experiment, we wanted to show that classification based on the top-n percepts, will improve classification accuracy dramatically, therefore providing evidence for our main hypothesis: There is no data like less data. For this experiments, we selected only training videos that contain a) the top-20 percepts (as determined by TFIDF), b) 20 random percepts and c) the low-20 percepts (as determined by TFIDF). We trained the SVM classification system overviewed in Section 7 (and detailed in [8, 4]) using the three options and measured the classification accuracy. Table 2 shows the results.

Even though we only selected on the level of videos rather than percepts and so not all ambiguities and noise has been filtered, the classification accuracy changes dramatically based on the selection of top percepts. Please note, that the numbers are not comparable with related work because the system was not tuned in any way as the only goal was to prove our theory. Also note that the results are based exclusively on audio and averaged over all five concepts/events.

We performed an analysis of the top-n percepts for each of the five concepts. We spot-listen the top-3 percepts by relating them to their corresponding segment in the audio file. For E001 Board tricks, music was the most common percept. For E002 Feeding an Animal the most common were silences, and speech. For E003 Landing a Fish the most common were silences and speech too. This is one of the reasons why the system was having troubles classifying these two classes. For E004 Wedding the most common were speech and music. Finally, for E005 Woodworking speech and tool noises were the most common.

Error	Baseline	Top 20	Low 20
False Alarm	6 %	6 %	6 %
Miss	72 %	66 %	79 %
EER	31 %	31 %	35 %

Table 2: Change of classification error using FA/Miss as defined by TRECVID MED and using Equal-Error-Rate for no percepts selection, top-20 frequent percepts and low-20 frequent percepts. Even though our method does not yet account for ambiguity, less data is better than all data.

9. CONCLUSION AND FUTURE WORK

We presented a theoretical framework that allows us to reason quantitatively and qualitatively about video concept detection even when multimedia documents have no obvious structure. Our framework is media independent and our supporting experiments were conducted on a publicly available large set of consumer-produced videos. Future work includes extending the framework and learning more from the natural language community and the original TREC evaluations (that later resulted in TRECVID).

Our next steps will also include building an actual system that will utilize the predictive power of our framework in different modalities by eliminating noisy percepts. A limitation of the approach is that percepts do not necessarily overlap with human percepts, therefore, introspection and analysis of concept detection results by human requires an intermittent step: a translator from machine-generated percepts-concept mappings, i.e. languages, to human languages. Another point of interest, specially when dealing with more than five events, is to remove speech and music from the audio. Then analyse what would be the most representative non-speech audio classes for each concept/event and compute the classification results.

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of 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 conclusion 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.

10. REFERENCES

[1] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[2] Sourish Chaudhuri, Mark Harvilla, and Bhiksha Raj. Unsupervised learning of acoustic unit descriptors for audio content representation and classification. In *INTERSPEECH*, pages 2265–2268, 2011.

[3] Alexander Hauptmann, Rong Yan, and Wei-Hao Lin. How many high-level concepts will fill the semantic gap in news video retrieval? In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pages 627–634, New York, NY, USA, 2007. ACM.

[4] Po-Sen Huang, Robert Mertens, Ajay Divakaran, Gerald Friedland, and Mark Hasegawa-Johnson. How to Put it into Words – Using Random Forests to Extract Symbol Level Descriptions from Audio Content for Concept Detection. In *In Proceedings of IEEE ICASSP*, pages AASP–P8.10, March 2012.

[5] Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Subhabrata Bhattacharya, Dan Ellis, Mubarak Shah, and Shih-Fu Chang. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, Gaithersburg, MD, November 2010.

[6] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl*, 2:1–19, 2006.

[7] Lie Lu and A. Hanjalic. Audio keywords discovery for text-like audio content analysis and retrieval. *Multimedia, IEEE Transactions on*, 10(1):74–85, jan. 2008.

[8] Robert Mertens, Po-Sen Huang, Gerald Friedland, and Ajay Divakaran. On the applicability of speaker diarization to audio indexing of non-speech and mixed non-speech/speech video soundtracks. *International Journal of Multimedia Data Engineering and Management*, to appear.

[9] Robert Mertens, Howard Lei, Luke Gottlieb, Gerald Friedland, and Ajay Divakaran. Acoustic super models for large scale video event detection. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, J-MRE '11, pages 19–24, New York, NY, USA, 2011. ACM.

[10] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso. On the use of audio events for improving video scene segmentation. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*, pages 1 –4, april 2010.

[11] Cees G. M. Snoek and Marcel Worring. Concept-based video retrieval. In *Foundations and Trends in Information Retrieval*, 2009.