Exploiting Temporal Sequence Structure for Semantic Analysis of Multimedia

Sourish Chaudhuri, Rita Singh, Bhiksha Raj

Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
{ SOURISHC, RSINGH, BHIKSHA}@CS.CMU.EDU

Abstract

Automatic deduction of semantic labels for audiovisual data requires awareness of context, which in turn requires processing sequences of audiovisual scenes or events. The representation of such sequences is important for semantic analysis tasks. Whereas, conventionally, sequences of specific short-duration event labels, often hand-annotated for learning detectors or classifiers, have been used, we propose a new technique for audiovisual event categorization in this paper, wherein units of audio and image scenes are discovered automatically from data in a likelihood-maximization process. We show how these units for audio and video, respectively called AUDs and VIDs, can be used to learn the salient characteristics of broad-category semantic labels without requiring explicit error recovery measures. Experiments with the MED-11 dataset show that AUDs and VIDs are better able to retrieve semantic categories from mixed-content data as compared to vector quantization-based systems and systems that use library-based descriptors.

Index: multimedia analysis, semantic labels, unsupervised lexicon learning, audiovisual data retrieval

1. Introduction

Indexing multimedia content, specifically audiovisual data content, for retrieval and other purposes has been a focus of attention for more than a decade. Early approaches to indexing audio data used speech recognition techniques to generate keywords for indexing [1]. Subsequently, approaches were developed for detecting specific sounds in audio streams such as gunshots, laughter, music, crowd sounds etc. [2] as well as to map words to acoustic phenomenon [3] using known vocabularies of sounds. Other analyses detect repeated sequences in television broadcast streams [4], to identify jingles and advertisements. Similarly, state-of-the-art event detection approaches in video employ lower level features (SIFT, GIST), texture information, spatio-temporal features, as well as libraries of semantic concepts, with detectors covering scenes, people, and various object types for event detection [5].

Approaches based on the use of libraries of concepts for training detectors are limited in their ability to represent data from new domains, and depend heavily on the semantic concepts in the library being relevant to the domain. Further, such libraries of concepts have typically been defined manually, and trained with data captured in clean, low-noise environments. As a result, the performance of detectors trained from these libraries may suffer when applied to noisy data, such as user-generated content on Youtube.

In this paper, we show that we can learn a lexicon of concepts for both the audio and video modalities in an unsupervised fashion. To capture real-world events and their various manifestations, the number of concepts for audio or video would be infinite. We reduce this set to a manageable size, to make unsupervised learning of concept lexicons from the available data possible, based on two hypotheses—first, even though the set of concepts in the real world may be infinite, a far smaller number occur in a dataset. Second, even if we use a smaller number of generalized concepts, we expect patterns of these units in the data to be characteristic of the larger events in the audio. Owing to their generalized nature, the learnt concepts do not always have clear semantic interpretations, but the representation of audiovisual data as a sequence of these concepts proves effective for automatic semantic analysis tasks. Further, we show that even though the process of transcription of data in terms of these concepts is noisy, they are sufficient to enable effective retrieval. We also show how the sequentaility of these concepts in the data can be exploited in different semantic contexts.

We leverage previous work [6] in automatically discovering units of sound, and describe an extension that allows application to both audio and video data, without requiring any kind of category-specific information at the time of lexicon learning. Unlike previous approaches that use Vector-Quantization (VQ) methods to create a discrete representation of a recording, we model multimedia data as being composed of a sequence of atomic units (analogous to concepts in a library), and our atomic units span a set of consecutive frames of variable length (similar to phonemes for speech). The intuition is that an observed event likely manifests over multiple frames, and identifying the sequence of frames as a block allows us to preserve information about the frequency of occurrence of the concept.

Our approach to semantic analysis using these lexicons is based on the observation that even though humans might recognize individual scenes, they make semantic associations based on sequences of such scenes; e.g. is a dull metallic sound caused by a hammer striking an object, a baseball bat hitting a ball, or a vehicle collision? The sound alone doesn’t provide us with sufficient information to infer the semantic context. However, if it is followed by applause, we may guess the context to be baseball, shrieks suggest an accident, while periodic repetitions of the metallic sound might suggest someone using a hammer. Our goal in our feature space design using atomic unit sequences for audiovisual data is to discover these “semantically meaningful” sequences. Results show that our assumptions and hypotheses about the use of these automatically discovered atomic units in semantic interpretation are effective, and improve significantly over VQ and concept-library-based baselines.

The rest of this paper is organized as follows: in Section 2, we discuss our approach to unsupervised learning of lexical units from audio and video data, while specifically noting the differences of our approach with class-specific Gaussian Mixture Models (GMM) and Hidden Markov Model (HMM) approaches. Section 3 describes one version of a semantic interpretation task. Section 4 describes our experimental setup and results, and conclude in Section 5.
2. Unsupervised Lexicon Learning

In previous work [6], we introduced automatically discovered units of sound, called Acoustic Unit Descriptors (AUDs). AUDs can be viewed as fundamental units that typically span a number of frames (unlike VQ). Since AUDs represent automatically discovered generalized units, they may not have a human-perceptible interpretation. In this paper, we generalize our formulation and introduce counterpart units for image, which we call Video Descriptors (VIDs). Similar to AUDs, VIDs use the intuition that semantic visual events are typically composed of a sequence of visual concepts, e.g. pitch throwing followed by hitting in baseball, and we would like to learn such concepts to create the lexicon of VIDs. Like AUDs, VIDs may span multiple successive frames in the video. Similar approaches have been used in the past to learn fundamental unit inventories for speech [7, 8], where discovered units were generally desired to be meaningful units of speech. We do not impose any such constraints and the automatically learned units are expected to be semantically polysemous. This gives us a great advantage in morphing to different levels of semantics when needed.

In past work, the process of learning of AUDs assumed that the AUDs are shared by all the classes, but employs class-specific language models (LM, henceforth), using the intuition that they would provide enhanced ability to distinguish between categories. In order to create a generalized lexicon learning framework, we relax these assumptions further in this paper, instead using a class-independent LM, which has 3 advantages: first, the lexicon learning phase is now entirely unsupervised, and no label information is required. This distinguishes our approach from standard GMM and HMM-based approaches, that train category specific models. Second, since we no longer need to decode the same test file for each candidate category using its native classifier. Features extracted from these transcriptions are then used to characterize the data at different semantic levels through a query, for example). Note that Eqn 1 in the lexicon learning algorithm allows us to obtain a transcription for any recording, in terms of the AUDs and VIDs as an abstraction when describing patterns in audio-visual data, we analyze LU sequences in order to characterize specific events or categories (which may have been generated through a query, for example). Note that Eqn 1 in the lexicon learning algorithm allows us to obtain a transcription for any recording, in terms of the LUs. By representing data using features derived from the sequences of decoded LUs, we enable the use of these features for training discriminative classifiers.

We now describe an iterative learning algorithm that allows us to obtain such a lexicon. Since AUDs and VIDs are synthetic concepts, we do not expect to have transcriptions of data in terms of these units, resulting in a fundamentally unsupervised approach to learning the lexicons. We employ a hill-climbing approach that learns parameters to maximize the likelihood of the entire observed training data. The iterative update mechanism used to learn the lexicon, is shown in Algorithm 1, and uses parameter estimates from iteration \( r \) to update the estimates in iteration \( r + 1 \). \( T \) refers to the best inferred transcript for a recording \( D_i \). \( A \) represent the AUD or VID parameters, and \( H \) represents the class-independent LM. Our learning paradigm is generative, and no discriminative class-specific information is used at the time of learning lexicons. We refer the reader to [6] for more details.

Depending on the task, one may need to use appropriate feature representations for the video. Since our task is semantically defined, our experiments in this paper use a semantically motivated feature set. We extract information from video frames using various object detectors that detect the number of people, faces, hands, vehicles, tires, crouching poses, wood and fur textures, and use these as a feature representation for each frame, which are then used to learn VIDs. Since these features are discrete integer valued, we transform them to a continuous space, using the process illustrated in Algorithm 2. This is done since continuous spaces are no less expressive than discrete ones, with the continuous space permitting more operations, if required. We chose to model the transformed space with gaussians, although any other appropriate density functions may be employed.

Since AUDs and VIDs are analogous in many ways, we use the term Lexical Units (LUs) as an abstraction when describing the use of these features for training discriminative classifiers.

We now describe an iterative learning algorithm that allows us to obtain such a lexicon. Since AUDs and VIDs are synthetic concepts, we do not expect to have transcriptions of data in terms of these units, resulting in a fundamentally unsupervised approach to learning the lexicons. We employ a hill-climbing approach that learns parameters to maximize the likelihood of the entire observed training data. The iterative update mechanism used to learn the lexicon, is shown in Algorithm 1, and uses parameter estimates from iteration \( r \) to update the estimates in iteration \( r + 1 \). \( T \) refers to the best inferred transcript for a recording \( D_i \). \( A \) represent the AUD or VID parameters, and \( H \) represents the class-independent LM. Our learning paradigm is generative, and no discriminative class-specific information is used at the time of learning lexicons. We refer the reader to [6] for more details.

Algorithm 1 Algorithm for learning LUs

\[
T_r^{i+1} = \arg\max_T P(T|D_i; H^r; A^r) \quad (1)
\]

\[
A^{r+1} = \arg\max_A \prod_{D_i} P(D_i|T_r^{i+1}; A) \quad (2)
\]

\[
H^{r+1} = \arg\max_H \prod_{D_i} P(T_r^{i+1}; H) \quad (3)
\]

Algorithm 2 d-dim discrete space \( \rightarrow \) continuous space

\[
D \leftarrow d\text{-dimensional discrete representation (n datapts x d)}
\]

\[
P \leftarrow PCA(D, n); P \text{ represents the d principal components}
\]

for \( i = 1 \rightarrow n \) do

\[
F_i \leftarrow \text{Component Loadings for } D_i
\]

end for

for \( j = 1 \rightarrow d \) do

Let continuous distribution of choice be \( N(\mu_j, \sigma_j^2) \)

\[
G_j \leftarrow CDF(N(\mu_j, \sigma_j^2)); \quad C_j \leftarrow CDF(F(:, j)) \quad (4)
\]

\[
T \leftarrow \text{Function to transform } C_j \text{ to } G_j
\]

\[
F(:, j) \leftarrow T(F(:, j)) \quad (5)
\]

end for

\[
P \leftarrow PCA(F, n); P \text{ represents the d principal components}
\]

for \( i = 1 \rightarrow n \) do

\[
F_i \leftarrow \text{Component Loadings for } P_i
\]

end for

\[
F \leftarrow \text{Continuous representation for further processing}
\]

Since continuous spaces are no less expressive than discrete ones, with the continuous space permitting more operations, if required. We chose to model the transformed space with gaussians, although any other appropriate density functions may be employed.

Since AUDs and VIDs are analogous in many ways, we use the term Lexical Units (LUs) as an abstraction when describing techniques that apply for both audio and video lexicons.

3. Semantic categorization with LUs

For the discovery of semantically meaningful patterns in audio-visual data, we analyze LU sequences in order to characterize specific events or categories (which may have been generated through a query, for example). Note that Eqn 1 in the lexicon learning algorithm allows us to obtain a transcription for any recording, in terms of the LUs. By representing data using features derived from the sequences of decoded LUs, we enable the use of these features for training discriminative classifiers.

The process described so far has not used any category specific information. Further, we can now use features derived from the LU sequences along with labels as input to a discriminative classifier. Features extracted from these transcriptions are then used to characterize the data at different semantic levels (in this paper, simply at the broad-category level), making this approach more flexible. In contrast, category-specific GMM or HMM based approaches model the entire concept class directly. While we do use HMMs in our work, we use them at an entirely different level: we model each LU as an HMM.

We evaluated the learnt LU lexicons on an event-detection task, where given a large dataset of recordings, we are required to retrieve all instances of a particular event in the dataset.
We used the Multimedia Event Detection (MED) dataset from TRECVID, 2011. The database consists of 15 event categories [9], with 2.4 K files belonging to these events, and 10K other recordings. We learnt a set of AUDs and VIDs separately from the same multimedia data files.

For each class $i$, where $i \in [1, 15]$, positive instances for that class are taken along with all the other files (negative instances for class $i$) and feature vectors are extracted from the decoded LU sequences for each of these instances, and a detector is trained for that class. We will discuss the effect of different feature sets in Section 4. At test time, for each instance in the test data, the detector for class $i$ predicts whether or not the test data belong to the class $i$. Performance can be gauged using Missed Detection (MD) and False Alarm (FA) rates, which are defined according to NIST guidelines for the MED11 task as follows: suppose there are $N_i$ test files, with $C_i$ belonging to class $i$, and $D_i$ of these were correct. Then:

$$MD = \frac{C_i - D_i}{C_i} \quad FA = \frac{N_i - D_i}{N_i - C_i}$$

(4)

Performance can be gauged using the Area Under ROC Curves (AUC), which measures MD rate against FA rate. Since both missed detections and false alarms are measures of error, the lower the area under the curve, the better the performance.

4. Experiments

For experiments using LU lexicons, each LU was modeled as a 5-state HMM, with a 4-Gaussian mixture state-output probability distribution and a class-independent unigram model over the LUs. Mel-frequency cepstra (MFCC) with deltas and double deltas were used for learning AUDs, while counts of various objects (discussed in Section 2) obtained from object detectors, subsequently transformed to a continuous form as described above, were used as features for learning VIDs.

First, we extracted unigram frequencies of various LUs, using the intuition that different categories would contain different LU distributions. We refer to these {LU, LUFrequency} feature sets as AUD–COUNT and VID–COUNT respectively. We also extracted a binary feature vector that indicates whether or not the LU occurred in the transcription of the data (called AUD–BINARY and VID–BINARY). In Section 4.3, we present results comparing these feature sets to baseline systems discussed in Section 4.1, as well as using enhanced feature sets with n-gram information to explicitly model LU sequences.

4.1. Baseline Systems

We compare performance of LU-based systems against VQ-based systems for both audio and video. VQ baselines were trained using various numbers of clusters and the best results are used for comparison. For audio, we set up two other baselines using semantically derived units. One was based on phoneme models learned from the HUB4 corpus [10]. The other baseline used the Art of Foley Sound Effects Library [11], which consisted of 480 individual events. Using a lexicon of audio units, each being one of these events, we transcribed the audio data. A classifier was then trained using count-frequency vectors of these units, as in the case of LUs. We refer to these systems as PHONE and FOLEY, respectively.

4.2. Random Forest Classifier

In order to predict whether a recording belongs to a particular class, we train binary classifiers for each category in a one-versus-all setting, using the random forest classifier [12]. Random forests are an extension of decision-tree classifiers, where the training process builds multiple trees instead of a single one. Given new test data, each of the trees in the forest returns a class label, which are used in a weighted vote to determine the predicted label. We chose the random forest classifier specifically because it has been shown to be resistant to overfitting. Trees in the forest were grown as far as possible, without pruning.

4.3. Results

Event-detection performance was measured using Area Under ROC Curves (AUC) based on MD and FA rates. Due to space constraints, we only discuss a few representative event categories here, but average results on all 15 categories are given in Table 1. First, we compare performance of the various categories using AUC—Fig 1 shows results on 3 categories; feeding an animal (most improvement over the nearest baseline) and woodworking project (least improvement), and fishing (median improvement). The following observations were noted across all 15 categories: a) Best overall performance is either the AUD-Count or VID-Count system, demonstrating that adapting the lexicon to the data rather than using a static lexicon is beneficial. b) AUD-Count setting outperforms all of VQ, Phone and Foley-based classifiers, and c) the binary LU-based classifier performs surprisingly well overall, outperforming the VQ baseline for video in the majority of categories, and performing comparatively to the Foley-based units for audio.

The results in Figure 1 were obtained using a lexicon size of 64. Realistically, the number of different semantic events in the data might be expected to be much larger than 64. Consequently, many real audio (video) events likely map to the same AUD (or VID). Fig. 2 shows the effect of increasing the lexicon size on performance for the previous three categories. Increasing the size of the AUD lexicon appears to improve performance for all categories, whereas the effect on VIDs is unclear.

While unigram characterizations are robust to noisy decodings, we may still expect higher order n-gram features to perform better since they provide more structure. However, we found performance with bigram and trigram characterizations to be worse than unigrams. This is likely because the process of decoding the data using the LUs is noisy, since multiple real-world events map to the same LU, as well as confusions due to

![Figure 1: AUC (y-axis) for various feature sets (lower is better)](image)

<table>
<thead>
<tr>
<th>Modality</th>
<th>LU-binary</th>
<th>LU-count</th>
<th>VQ</th>
<th>Foley</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio AUC</td>
<td>0.279</td>
<td>0.227</td>
<td>0.291</td>
<td>0.273</td>
<td>0.262</td>
</tr>
<tr>
<td>Video AUC</td>
<td>0.259</td>
<td>0.244</td>
<td>0.283</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
noise in the recordings. Further, recordings in the same semantic category often vary greatly in how the semantics are conveyed, rendering simple bigram characterizations ineffective. To counter this, we used flexible $n$-grams, where for a given LU, we look at all the LUs within a context window, and increment the count of a flexible $n$-gram, when the context LUs all appear in the context window. We expect this to reduce the effect of noisy decodings and sequence variations. Figure 3 compares performance of the unigram, bigram and flexible bigram characterizations for using a 64-AUD lexicon, across five categories. We found that a window length of 6 performed best for flexible bigrams, followed by 3.

For the MED11 task, NIST provides benchmark operating points which was 75% missed detection at 6% false alarm rate. Figure 4 shows the complete ROC plots for 5 categories with the red circle being the benchmark operating point. We find that our LU-based approach performs better than the benchmark for all categories, and is well clear of the benchmark for a number of categories (e.g. flash mob, repairing appliance).

5. Conclusions

Our results show that data-driven lexicons with no guaranteed semantic import outperform semantic concept libraries and VQ-based approaches significantly. Limited sets of semantically-defined units (e.g. Foley set) cannot adequately describe all events that occur in even a small collection of recordings. Further such libraries are often created with studio-quality recordings and may be ineffective in analyzing lower quality, user-generated content. As a result, data-driven units provide a distinct advantage in learning a vocabulary to fit the data without making significant prior assumptions, and this is borne out by our experiments. The performance can be improved further by combining audio and video features using fusion techniques, which we expect to explore in the future.

The units obtained through data-driven discovery are not entirely lacking in semantic association. The performance of the LU-BINARY systems shows that merely detecting presence of LUs performs significantly better than random, probably due to capturing some underlying semantic. Moreover, the LUs themselves are generatively learnt without any explicit requirement to be discriminative, yet perform well on a discriminative task, further strengthening the notion that they capture some underlying semantic in the data.

6. Acknowledgements

This work was supported by funding from Charles Stark Draper Labs, Boston, USA. Video detectors for the VIDs were provided by Intuvision Inc., Boston, USA.

7. References