

# Audio-Visual Event Classification via Spatial-Temporal-Audio Words

Yu Cao, Sung Baang, Shih-Hsi “Alex” Liu, Ming Li, Sanqing Hu\*

*Department of Computer Science, California State University, Fresno, CA, USA*  
{yucao,sibang2002,shliu, mingli}@csufresno.edu

*\*Department of Neurology, Mayo Clinic, Rochester, MN, USA*  
{Hu.Sanqing}@mayo.edu

## Abstract

*In this paper, we propose a generative model-based approach for audio-visual event classification. This approach is based on a new unsupervised learning method using an extended probabilistic Latent Semantic Analysis (pLSA) model. We represent each video clip as a collection of spatial-temporal-audio words, which are generated by fusing the visual and audio features using the pLSA model. Each audio-visual event class is treated as the latent topic in this model. The probability distributions of the spatial-temporal-audio words are learnt from training examples, which include a sequence of videos that represent different types of audio-visual events. Experimental results show the effectiveness of the proposed approach.*

## 1. Introduction

Categorizing the video sequence into audio-visual event class is an important field of content-based video analysis research. The goal of audio-visual event categorization is to automatically classify a given video clip into different sets of events, such as fighting, singing, and drinking. The ability of categorizing audio-visual data is very useful for a wide range of applications.

Many of the existing methods only employ visual information for event categorization. Aggarwal and Cai [1] provide comprehensive reviews for historical work. Some popular earlier work includes feature or tracking based approach [2]; shape-based approach [3]; flow-based methods [4]; and methods based on space-time interest points [5, 6]. Another avenue for event

classification is to use both the visual and audio information. These efforts are particularly plausible for specific domains, such as sports videos [7], and surveillance videos [8]. Sadlier et al. [7] explore the solution of employing both audio and visual features by Support Vector Machine (SVM). Cristani et al. [8] propose a method based on “Audio-video concurrence” matrix to integrate the audio and visual information for scene analysis for surveillance video. Central to all these work are models that are very effective for the particular domain. However, they may lack the capacity to handle other video genres.

In this paper, we propose a new method to integrate the visual and audio information for video event categorization. Our method is motivated by recent efforts on three aspects: space-time interest points [5, 9], object class detection [10, 11], and human action categorization based on probabilistic graphical models [12]. In those papers [10-12], generative graphical models, such as probability Latent Semantic Analysis (pLSA) [13] and Latent Dirichlet Allocation (LDA) [14], are used to learn and recognize the object or the human action event. However, those papers only consider the visual information for event classification. In our method, we extend the pLSA model to incorporate both visual and audio information. The major contribution of this work is the new representation of a video sequence by spatial-temporal-audio “words”. These “words” are generated from the visual and audio feature descriptors using the extended pLSA model. The new representation enables the unsupervised learning of audio-visual event classification. Our method works well for different types of videos, such as entertainment videos (e.g., fighting or shouting scenes in Hollywood films), home videos (e.g., riot event in news videos), and etc.

The rest of this paper is organized as follows. Section 2 introduces the structures of the extended pLSA model, as well as the learning and recognition procedures. Section 3 presents the performance of our methods. Finally, we offer our concluding remarks in Section 4.

## 2. Proposed Approach

Figure 1 depicts an overview of our method. There are two components in our system. The first one is “Learning Component” and the second is “Recognition Component.” The “Learning Component” is shown in the left side of the figure. The goal of this part is to build the model. The input of this component includes a sequence of videos for each category. Our algorithms can generate a model for each category. For the “Recognition Component” shown in the right side of the figure, our method takes a new video clip with unknown category as input. The models trained from the “Learning Component” are used to classify the unknown video sequence into one of the categories. In the following sub-sections, we first introduce the structure of the extended pLSA model. Then we present our algorithms for learning the model parameters, followed by a detailed description for recognizing an instance of an unknown video clip.

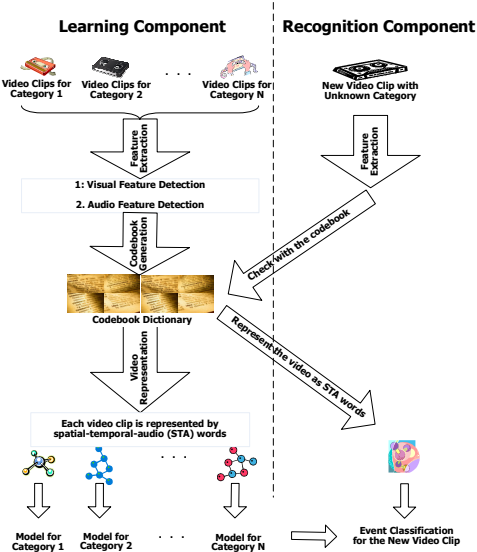


Figure 1: System overview of the proposed audio-visual event categorization algorithm using spatial-temporal-audio words.

### 2.1. Structure of the Extended pLSA Model

We use an extended pLSA model to encode the visual and audio information for each video. The pLSA model [13] is a statistical technique for the analysis of two-mode and co-occurrence data with a wide range of

applications in information retrieval and natural language processing. It is recently employed by computer vision community for solving the problems of object class recognition [10, 11] and human action categorization [12]. However, this simple “bag of words” model used in [10-12] can not be directly applied to incorporate audio information. We extend the pLSA model by adding a new random variable, which is used to represent the audio features. In the following descriptions, we present the models follow the terms and conventions introduced in [10-13].

Suppose we have  $D$  ( $D = \{d_1, \dots, d_N\}$ ) video clips where  $d_i$  represents the  $i^{\text{th}}$  video that contains both audio and visual information. We use  $w_{S-T}$  to represent the spatial-temporal words, which are similar to the concepts proposed in [12]. To incorporate audio information, we extend the pLSA model by adding a new random variable  $w_A$  to represent the audio features. Assume the spatial-temporal vocabulary is represented as  $W_{S-T} = \{w_{S-T-1}, \dots, w_{S-T-M}\}$  while the audio vocabulary is  $W_A = \{w_{A-1}, \dots, w_{A-K}\}$ . The corpus of the videos can be summarized in an three-dimensional co-occurrence matrix  $\bar{N}$ , whose degree is  $M \times K \times N$  and the entry  $n(w_{S-T-m}, w_{A-k}, d_n)$  in this matrix represents how often the term  $w_{S-T-m}$  and  $w_{A-k}$  occurred in video  $d_n$ . A latent topic variable  $z$  is used to associate the occurrence of word  $w_{S-T}$  and  $w_A$  to video  $d$ . In our context, this latent variable indicates the audio-visual event category. The joint probability model over  $W_{S-T} \times W_A \times D$  is represented by the following equation:

$$P(w_{S-T}, w_A, d) = P(d) \cdot P(w_{S-T}, w_A | d). \quad (1)$$

From Equation (1), we can perform further derivation by importing the latent variable  $z$ .

$$P(w_{S-T}, w_A, d) = \sum_{z \in Z} P(z) P(d | z) P(w_{S-T}, w_A | z). \quad (2)$$

We use the Expectation-Maximization (EM) algorithm for both training and testing. EM alternates two steps: (1) an expectation (E) step where posterior probabilities are computed for the latent variables, (2) an maximization (M) step, where parameters are updated. In our extended pLSA model, the E-step equation is listed as below.

$$P(z | w_{S-T}, w_A, d) = \frac{P(z) P(d | z) P(w_{S-T}, w_A | z)}{\sum_{z' \in Z} P(z') P(d | z') P(w_{S-T}, w_A | z')}. \quad (3)$$

The formulas for the M-step are listed as follows

$$P(w_{S-T}, w_A | z) \propto \sum_{d \in D} n(w_{S-T}, w_A, d) P(z | w_{S-T}, w_A, d). \quad (4)$$

$$P(d | z) \propto \sum_{w_{S-T} \in W_{S-T}} \sum_{w_A \in W_A} n(w_{S-T}, w_A, d) P(z | w_{S-T}, w_A, d). \quad (5)$$

$$P(z) \propto \sum_{d \in D} \sum_{w_{S-T} \in W_{S-T}} \sum_{w_A \in W_A} n(w_{S-T}, w_A, d) P(z | w_{S-T}, w_A, d). \quad (6)$$

A new video  $d^*$  is classified by executing the EM with  $P(w_{S-T}, w_A | z)$  fixed and computing the  $P(z | d^*)$  to determine the event category. See the Section 2.3 (“Recognition”) for details on classifying a new video.

## 2.2. Learning

There are two steps in the learning stage: visual and audio feature extraction, and spatial-temporal-audio “words” generation. The goal is to determine the distribution of the spatial-temporal-audio “words” over the topic.

In the first step (feature extraction), there are two components: visual feature extraction and audio feature extraction. To obtain the visual features, we use the space-time interest point detector technique, which is similar to the one introduced in [5]. Separable linear filters are used in this method. They are applied to the video to obtain the response function. The space-time interest points are extracted around the local maxima of the response function. At each interest point, a cuboid is extracted and it contains the spatio-temporally windowed pixel values. The size of the cuboid is selected to contain the majority of the volume of data that contribute to the response function at that interest point. Feature descriptors are extracted for each spatial-temporal cuboid. We use the brightness gradient as the feature descriptors and apply PCA to reduce the dimensionality of the descriptors. To generate the audio feature descriptors, we use a windowed scan over the audio stream. The windowed scan includes sliding a window over the audio data in fixed increments. Audio features are extracted from each sliding window using short-time Fourier transform (STFT) techniques. The reason to choose STFT is because STFT provides a rich representation that is capable of modeling a variety of perceptual characteristics such as pitch and loudness and non-perceptual features such as the approximate band-width of the audio [15].

After the feature extraction step, we need to combine the visual and audio feature descriptors to generate the spatial-temporal-audio “words”. We first generate the codebooks by applying k-means clustering to both the visual feature descriptors and the audio feature

descriptors. From these codebooks, we can vector-quantize the visual and audio feature descriptors into spatial-temporal-audio “words”. Each video clip is represented by a two-dimensional matrix that indicates the co-occurrence of the spatial-temporal-audio “words” in this video. Therefore the entire training data, which is a sequence of video clips, is represented by a three-dimensional matrix. We apply the EM algorithm to this three-dimensional co-occurrence table and obtain the model parameters.

## 2.3. Recognition

The goal of the recognition component is to determine the category of a given new video. The first step for recognition is to extract the visual and audio feature descriptors from the input video. Based on these descriptors and the codebooks (which is generated during the learning stage), we could “project” the new video on the simplex spanned by the  $P(w_{S-T}, w_A | z)$ , which is the spatial-temporal-audio “words” distribution over a latent topic. Given a new video clip  $d_{test}$  with unknown category, the correct category  $z_k$  of this video should satisfy the following equation:  $z_k = \underset{z_k \in \mathcal{Z}}{\operatorname{argmax}} P(z_k | d_{test})$ . To calculate  $p(z_k | d_{test})$ ,

we apply Bayes rule to generate the following equation:

$$P(z_k | d_{test}) = \frac{P(d_{test} | z_k) \cdot P(z_k)}{P(d_{test})}. \quad (7)$$

In order to obtain the likelihood and the prior in Equation (7), an EM algorithm that is similar to the one used in training stage can be employed. Different from the EM method for training, the value of  $P(w_{S-T}, w_A | z)$  is fixed during the EM execution and this value is obtained from the learning stage.

## 3. Experimental Results

The proposed algorithms were implemented using Matlab. To show the effectiveness of our approach, we select different human activities extracted from a diverse range of sources. Five visual-audio event categories from movies and home videos are used: crying, shouting, laughing, drinking, and eating. For each category, we use ten videos, three hundred frames each. The frame rate is thirty frames per second and the average length of each video is about ten seconds. We build the codebook from two videos of each category. For the remaining forty videos (eight videos for each of the five categories), cross-validation is selected to test the performance of the algorithm. For each run, we train the model using thirty five videos (seven videos from each of the five categories) and test the model

from the remaining five videos (one video per category). Please note, the ten videos (two videos from each of the five categories) used for generating the codebooks are not re-used for the cross-validation. Table 1 shows the confusion matrix for our dataset. The overall performance is encouraging with average accuracy at 81%. However, the confusions between “crying” and “shouting”, as well as “drinking” and “eating”, is large. The main reason is because the visual and audio properties between “crying” and “shouting” category are similar to each other. The same reason is applied to the “drinking” and “eating” category. For performance comparison, we implement other event classification algorithms. The first compared algorithm, defined as algorithm A, uses similar visual feature extraction methods and unsupervised learning framework as our proposed approach, but does not utilize the audio features. The second compared algorithm, defined as algorithm B, only uses audio features. The average accuracy of algorithm A and algorithm B are 70% and 62% respectively. These experiments show that the proposed method is more effective because of the integration of both visual and audio features. If the testing videos do not contain detectable audio features, the proposed method will only utilize the visual information. In this case, the performance of our method is on par with existing

**Table I**  
**Confusion Matrix of the Proposed Approach**

|                 |             |             |             |             |             |
|-----------------|-------------|-------------|-------------|-------------|-------------|
| <b>crying</b>   | <b>0.75</b> | 0.16        | 0.09        | 0.00        | 0.00        |
| <b>shouting</b> | 0.11        | <b>0.81</b> | 0.08        | 0.00        | 0.00        |
| <b>laughing</b> | 0.03        | 0.03        | <b>0.91</b> | 0.01        | 0.02        |
| <b>drinking</b> | 0.00        | 0.00        | 0.00        | <b>0.77</b> | 0.23        |
| <b>eating</b>   | 0.00        | 0.00        | 0.00        | 0.19        | <b>0.81</b> |

visual information-based methods [12].

#### 4. Conclusions and Future Work

In this paper, we have demonstrated a new method to integrate the visual and audio information for the purpose of audio-visual event classification. An extended pLSA model is developed to fuse the visual and audio information. EM algorithm is used for both learning and testing stage. Experiments on different types of videos validated the proposed methods. In the near future, we plan to refine the current implementation and apply the algorithms to larger datasets. During our investigation, we noticed the lack of challenging and standardized datasets for thoroughly validation and performance comparison. This can be an interesting future topic. One thing we want to point out is that the focus of this paper is to classify the pre-segmented videos. To apply the proposed approach to the full-length video, we should first perform

shot/scene segmentation. It will be an interesting future topic to integrate the proposed approach with existing shot/scene segmentation techniques. Other topics for future investigation include improving the proposed methods to localize different audio-visual events simultaneously in a complex video sequence.

#### 5. References

- [1] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, 1999.
- [2] A. Yilmaz and M. Shah, "Recognizing Human Actions in Videos Acquired by Uncalibrated Moving Cameras," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, Beijing, P.R.China, 2005, pp. 150-157.
- [3] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, pp. 249 - 257, 2006.
- [4] E. Shechtman and M. Irani, "Space-Time Behavior-Based Correlation-OR-How to Tell If Two Underlying Motion Fields Are Similar Without Computing Them?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 2045 - 2056, 2007.
- [5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. of Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, Beijing, P.R.China, 2005, pp. 65-72.
- [6] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," in *Proc. of British Machine Vision Conference (BMVC)*, Edinburgh, U.K, 2008, pp. III:1249.
- [7] D. A. Sadlier and N. E. O'Connor, "Event detection in field sports video using audio-visual features and a support vector Machine," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 1225 - 1233, 2005.
- [8] M. Cristani, M. Bicego, and V. Murino, "Audio-Visual Event Recognition in Surveillance Video Sequences," *IEEE Transactions on Multimedia*, vol. 9, pp. 257 - 267, 2007.
- [9] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, Nice, France, 2003, pp. 432 - 439.
- [10] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering object categories in image collections," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, Beijing, P.R.China, 2005, pp. 370- 377.
- [11] R. Fergus, "Visual Object Category Recognition," in *Department of Engineering Science*, vol. Doctor of Philosophy. Oxford: Oxford University, 2005, pp. 1-193.
- [12] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," in *Proc. of British Machine Vision Conference (BMVC)*, Edinburgh, U.K, 2006, pp. III:1249.
- [13] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, vol. 42, pp. 177-196 2001.
- [14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993 - 1022, 2003.
- [15] D. Hoiem, Y. Ke, and R. Sukthankar, "SOLAR: sound object localization and retrieval in complex audio environments," in *Proc. of IEEE International Conference on Acoustics*,

*Speech, and Signal Processing (ICASSP)*, Philadelphia, PA,  
USA, 2005, pp. 429-432.