GRAPH BASED MULTIMODAL WORD CLUSTERING FOR VIDEO EVENT DETECTION

Aravind Vembu, Pradeep Natarajan, Shuang Wu, Rohit Prasad, Prem Natarajan

Speech, Language and Multimedia Business Unit,
Raytheon BBN Technologies, 10 Moulton Street, Cambridge, MA 02138
{avembu,pradeepn,swu,rprasad,pnataraj}@bbn.com

ABSTRACT

Combining diverse low-level features from multiple modalities has consistently improved performance over a range of video processing tasks, including event detection. In our work, we study graph based clustering techniques for integrating information from multiple modalities by identifying word clusters spread across the different modalities. We present different methods to identify word clusters including word similarity graph partitioning, word-video co-clustering and Latent Semantic Indexing and the impact of different metrics to quantify the co-occurrence of words. We present experimental results on a ≈45000 video dataset used in the TRECVID MED 11 evaluations. Our experiments show that multimodal features have consistent performance gains over the use of individual features. Further, word similarity graph construction using a complete graph representation consistently improves over partite graphs and early fusion based multimodal systems. Finally, we see additional performance gains by fusing multimodal features with individual features.

1. INTRODUCTION

With the explosion of video content over the internet, automatic understanding of videos is a task which has several applications including search, retrieval and large volume data management and storage. A key feature of video content is that it consists of different modalities, the correlations between which can be exploited for better understanding of the content. The nature of these correlations (and which of them can actually be detected) range from the direct, like a dog seen and heard barking in the video, to the subtle like the singing of a certain song during birthday parties.

Features for videos (and images) are commonly based on Bag of Words (BoW)[1] models which have shown good performance on many tasks including video event detection [2]. Here, low-level visual (e.g. [3][4]) and audio (e.g. [5]) features are extracted from videos, projected to a pre-trained codebook and then pooled to get a video-level feature representation. Traditionally, information from multiple such modalities are integrated using early fusion, such as feature concatenation and multiple kernel learning [6], or late fusion such as score level fusion [2].

An alternate approach is to discover multimodal features based on the individual unimodal features and their co-occurrences at the video [7] and temporal [8] levels. In [9] dependencies between the modalities are found through extensive processing during feature extraction and a vocabulary is built based on these dependencies. An alternate approach is to exploit the co-occurences at the video level [7] to discover a multimodal vocabulary based on individual unimodal video-level features seen in a training corpus. A key advantage of this method is that it can be applied on pre-extracted features from multiple modalities with only a small additional post-processing cost.

In this work, we take the approach from [7] and explore a large set of variations and extensions to multimodal vocabulary learning. We explore variations in the word similarity graph partitioning methods and compare similarity metrics for word co-occurrence and optimal graph construction choices between partite and complete graphs. We also evaluate alternate word clustering methods in addition to word similarity graph partitioning [7], such as word-document co-clustering [10] and latent semantic indexing [11]. We rigorously evaluate the performance of each of these methods for the video event detection task on the highly diverse and challenging TRECVID MED 2011 dataset [12].

The rest of the paper is organized as follows: section 2 describes the bag-of-words based video event detection system; section 3 describes the different codeword similarity metrics considered; section 4 details the various methods for discovering the multimodal words; section 5 presents experimental results; and section 6 describes our conclusions.

2. VIDEO EVENT DETECTION SYSTEM

For the video event detection task, we rely on three commonly used feature sets: SIFT [3] (visual features extracted at the frame level), STIP [4] (visual features over spatio-temporal volumes), and audio features (14 MFCC's [5] and audio energy over overlapping audio frames, along with their first and second derivatives) to capture information from distinct modalities. For each of these features, we compute a codebook by unsupervised clustering from a training set of

videos. Given a new video, we extract the low level features and project each feature extracted to the codebook using soft quantization [13]:

$$\alpha_{i,j} = \frac{\exp\left(-\beta||\boldsymbol{x}_i - \boldsymbol{c}_j||^2\right)}{\sum_{k=1}^{K}\exp(-\beta||\boldsymbol{x}_i - \boldsymbol{c}_k||^2)}, \quad (1)$$

where β controls the soft assignment. For pooling we take the average of the soft-assignments for each region, i.e. $\mathbf{h}_m = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\alpha}_i$. We train Support Vector Machine (SVM) classifiers for each event using χ^2 kernels computed as:

$$K(\mathbf{x}, \mathbf{y}) = e^{-\rho\sum_i \frac{(x_i - y_i)^2}{x_i + y_i}}. \quad (2)$$

During training, we perform an extensive grid search to determine the optimal value for the slackness penalty parameter C in the SVM and the parameter ρ of the kernel.

3. WORD SIMILARITY METRICS

To discover multimodal word clusters, we measure word similarity between codewords of pre-extracted features. These individual features are extracted from different modalities in the video and represented by a bag of words based histogram. For a given video, we normalize the features in each modality by the L_1 norm of the features to prevent a bias towards longer videos. These feature values are represented by $h(v, x)$, corresponding to the L_1 normalized histogram values for a given video v and codeword x. We describe in this section alternate measures of word similarity between the codewords of features in different modalities.

3.1. TF-IDF based Measure

Term frequency of word x in video v can thought of as represented by $h(v, x)$, and the document frequency of word x can be taken as the summation of $h(v, x)$ over all the training videos. This gives a TF-IDF type measure as defined in equation 3, where $v_1 \cdots v_N$ represent the N training videos.

$$\text{TF-IDF}(x, v) = \frac{h(v, x)}{\sum_{i=1}^{N} h(v_i, x)} \quad (3)$$

We define similarity between two words x and y as the linear kernel of the TF-IDF vector taken over all the training documents. This measure of similarity was used in [7].

3.2. Normalized Pointwise Mutual Information (NPMI)

Another measure of dependency between the video and word as suggested by Liu et.al. in [14] is the use of pointwise mutual information (PMI). Since PMI values can go to negative infinity, we use the Normalized PMI (NPMI) as given in equation 4 to constrain the PMI value to the range $[-1, 1]$.

$$\text{NPMI}(x, v) = \frac{\log\frac{p(x,v)}{p(v)p(x)}}{-\log p(x, v)} = \frac{\log\frac{p(x|v)}{p(x)}}{-\log p(x|v)p(v)} \quad (4)$$

While Liu et.al [14] treat $h(v, x)$ as an empirical joint probablity $p(v, x)$, we opt for the more commonly used interpretation of the bag of words histogram feature as a measure of the conditional probablity $p(x|v)$ (as in [15]). Assuming that the training videos are equally likely (i.e., $\forall N p(v_i) = 1/N$), we have an estimate of $p(x)$ as given by equation 5 and NPMI given by equation 6.

$$p(x) \approx \sum_{i=1}^{N} p(x|v_i) \cdot p(v_i) = \frac{1}{N}\sum_{i=1}^{N} h(v_i, x) \quad (5)$$

$$\text{NPMI}(x, v) \approx \frac{\log\frac{N \cdot h(v,x)}{\sum_{i=1}^{N} h(v_i,x)}}{-\log\frac{h(v,x)}{N}} \quad (6)$$

We define NPMI based similarity between codewords x and y using a radial basis function (RBF) kernel over the NPMI vectors as in [14].

4. MULTIMODAL WORD DISCOVERY

We discover multimodal word clusters that capture relationships between words (within the same modality or across modalities), and find the final feature vectors for the videos by a simple average pooling of the individual feature values of the words in the cluster. These new multimodal word based features are used to train a classifier for event detection as detailed in Section 2. We explore several methods for discovering the word clusters, detailed in this section.

4.1. Word Similarity Graph Partitioning

The idea of discovering word clusters using a word-word similarity matrix for videos was first proposed in Liu et.al. [14] where the different modalities were different views of the same action. This was adopted in [7] for finding bimodal words across visual and audio modalities. They construct a bipartite graph between the words of two different modalities and use a spectral graph partitioning technique to identify clusters on this graph.

The similarity between two words can be defined as suggested in Section 3. Once the graph similarity matrix has been constructed, spectral graph clustering algorithms such as normalized graph cut [16] can be used to identify word clusters. While [7] restricts itself to the use of a bipartite graph, we explore the use of partite graph constructions (bipartite for 2 modalities and multipartite for more) as well as complete graphs (where similarities between words in the same modality are also found and can be non-zero in the graph similarity matrix). The use of complete graphs enables a comparison with other traditional word clustering methods from natural language processing described in 4.2 and 4.3.

4.2. Word-video Co-clustering by Spectral Graph Partitioning

Dhillon et. al. [10] suggested a method to perform word-document co-clustering which is also based on the construction of a bipartite graph and the use of spectral graph partitioning. A word-document bipartite graph is constructed using simple TF or TF-IDF measures as similarity between the word and document nodes in the graph. We adapt the method and create a word-video bipartite graph using the TFs where the word may be from any of the unimodal vocabularies.

Given the word-video TF matrix, $\mathbf{A}$, of size $M \times N$ (number of words $\times$ number of training videos), we find $\mathbf{A_n} = \mathbf{D_1}^{-1/2}\mathbf{A}\mathbf{D_2}^{-1/2}$, where $\mathbf{D_1}$ and $\mathbf{D_2}$ are diagonal graph node degree matrices given by $D_1(i,i) = \sum_j A_{ij}$ and $D_2(j,j) = \sum_i A_{ij}$. Dhillon et. al. [10] show that the singular value decomposition (SVD) of $\mathbf{A_n}$ can provide a normalized cut of the word-document bipartite graph, which helps reduce computation. Specifically, to find k word clusters, we use the $l = \lceil log_2 k \rceil$ left singular vectors $\{\mathbf{u_2}, \ldots, \mathbf{u_{l+1}}\}$ and right singular vectors $\{\mathbf{v_2}, \ldots, \mathbf{v_{l+1}}\}$ of the SVD of $\mathbf{A_n} = \mathbf{U\Sigma V^T}$ to construct $\mathbf{Z}$ as follows

$$\mathbf{Z} = \begin{bmatrix} \mathbf{D_1}^{-1/2}\mathbf{U} \\ \mathbf{D_2}^{-1/2}\mathbf{V} \end{bmatrix} \tag{7}$$

where $\mathbf{U} = [\mathbf{u_2}, \cdots, \mathbf{u_{l+1}}]$ and $\mathbf{V} = [\mathbf{v_2}, \cdots, \mathbf{v_{l+1}}]$. We determine the word clusters on this l dimensional data $\mathbf{Z}$ using standard k-means clustering.

4.3. Latent Semantic Indexing

Another methodology we implement to identify word clusters is Latent Semantic Indexing (LSI) using the word-video TF matrix $\mathbf{A}$ as is done for documents [11]. This relies on a direct SVD of $\mathbf{A}$ into $\mathbf{U\Sigma V^T}$ with the l largest left singular vectors providing a low dimensional projection of the words. The lower dimensional representation of words are given by $\mathbf{U}_{Nxl} \cdot \mathbf{\Sigma}_{lxl}$ where the subscripts give the sizes of the matrices, and these may again be clustered using k-means clustering.

5. EXPERIMENTAL RESULTS

We test our approach on a large, benchmark dataset of videos used in the TRECVID MED 2011 evaluations [12]. For our experiments, we use the development set released for the evaluations. This is a 13,000 video set containing ~1,750 unconstrained web videos from 10 complex events of interest, with 100-200 examples per event, and the remainder from the background class. We partition the videos from each event and the background class in a 3:1 ratio into train and test sets. The training set is used to train the final SVMs, after tuning of the hyper parameters C and ρ (Section 2) by k-fold validation on the training set.

Table 1. Baseline video event detection using ANDC.

Feature	ANDC
SIFT	0.9938
STIP	0.8895
MFCC	1.043

We compare systems using the average normalized detection cost (ANDC) defined in [12]. For each event of interest, we first determine a detection threshold Th on the output probability score using k-fold validation. Then the ANDC score is computed as:

$$f = w_{MD}P_{MD}(Th) + w_{FA}P_{FA}(Th) \tag{8}$$

where $P_{MD}(Th)$ and $P_{FA}(Th)$ are the missed detection and false alarm rates at the detection threshold Th and w_{MD}, w_{FA} are the relative weights for missed detections and false alarms. In particular, a lower ANDC score indicates better performance. This score is often used in machine learning with imbalanced datasets. For the TRECVID MED dataset, different systems are compared with w_{MD}=1.0, w_{FA}=12.49.

5.1. Baseline Results

To create a baseline system, we use individual features to perform event detection. Table 1 gives ANDC results for SIFT, STIP and MFCC features.

5.2. Impact of Word Similarity Metric

We ran experiments to identify the better similarity metric between TF-IDF and NPMI for the word similarity graph partitioning method described in Section 4.1. All results in Table 2 are based on the usage of a partite graph and for varying number of word clusters. The given number reflects the maximum number of clusters provided as input to the normalized cut algorithm; the actual number of found clusters may be fewer. Results in Table 2 show that increasing the number of clusters results in improved performance. The better performance of the TF-IDF measure over NPMI in general, especially at higher number of clusters, can also be noted. Analysis of the discovered clusters show that the use of TF-IDF results in more reasonable clusters of moderate size, while the use of NPMI as a metric generates many singular clusters (clusters with only a single word) and clusters with a large number of words, resulting in a highly bimodal distribution of cluster sizes.

5.3. Comparison of Partite and Complete Graph

Moving from a partite graph to a complete one provides consistent performance gains, as detailed in Table 3, suggesting a better identification of clusters. The use of a complete graph results in discovery of more unimodal word clusters and fewer

Table 2. Comparison of similarity metrics and number of clusters using ANDC.

Max. no. of clusters	SIFT+MFCC		STIP+MFCC		SIFT+STIP+MFCC	
	PMI	TF-IDF	PMI	TF-IDF	PMI	TF-IDF
2000	1.1470	0.9460	0.8835	0.9063	0.8524	0.9461
4000	1.0070	0.9080	0.8543	0.9206	0.8500	0.8395
6000	0.9455	0.9079	0.8694	0.8987	0.8338	0.8292

Table 3. Multimodal event detection performance using ANDC.

Method	SIFT+ MFCC	STIP+ MFCC	SIFT+ STIP+ MFCC
Word sim + partite graph	0.9079	0.8987	0.8292
Word sim + complete graph (**SGfeat**)	0.8771	0.8632	0.8062
Word-video co-clustering (**CCfeat**)	0.9343	0.8371	0.8034
LSI (**LSIfeat**)	0.9043	0.8332	0.8089

multimodal words than with the use of partite graphs which resulted in unexpectedly large number of multimodal clusters (and almost no unimodal clusters). Results have been reported only for number of maximum clusters set at 6000, although similar trends were observed at lower number of clusters as well.

The use of a complete graph is akin to word clustering without consideration of modality, allowing a direct comparison with the word-video co-clustering and LSI methods. These sets of at most 6000 features formed from multimodal words are referred to as **SGfeat**, **CCfeat** and **LSIfeat** for those found by word Similarity (complete) Graph partitioning, word-video Co-Clustering and Latent Semantic Indexing respectively. Video event detection performance using any of **SGfeat**, **CCfeat** or **LSIfeat** outperforms the baseline use of unimodal features shown in Table 1.

5.4. Early Fusion with Individual Features

While all three considered methods provide gains in performance over individual features, the question is whether these multimodal word clusters capture some underlying patterns. Therefore, we performed early fusion of unimodal features with the multimodal word features using MKL [6] and compared the performance with the early fusion of only the unimodal features. The resulting performance improvements, as seen in Table 4, suggest that these features provide information complementary to unimodal features.

Table 4. MKL based early fusion performance using ANDC.

Features combined using MKL	ANDC
SIFT + STIP + MFCC	0.8288
SIFT + STIP + MFCC + SGfeat	0.8055
SIFT + STIP + MFCC + CCfeat	0.8055
SIFT + STIP + MFCC + LSIfeat	0.8081

6. CONCLUSIONS

We presented a rigorous analysis of graph-based clustering approaches for designing multimodal features. Our results indicate that TF-IDF with a linear kernel provides a better measure of co-occurrence of codewords over use of NPMI with RBF kernel, evident in the better performance of the former in discovering word clusters using the word similarity graph partitioning method. Use of complete graphs instead of partite graphs provides further gains. The discovery of multimodal words, using any of the suggested methods, and use of combined features based on these clearly helps in the event detection task over the use of unimodal features alone. The improvement in event detection over early fusion using the original unimodal features shows that multimodal features are capturing underlying multimodal patterns that aid in event detection. Additionally, while this entire method was unsupervised, in future we plan to discover event based codeword clusters to provide more discriminative information. Further, our approach can also be used to discover better word clusters based on word co-occurrence in temporal and spatial levels.

7. ACKNOWLEDGMENT

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC or the U.S. Government.

8. REFERENCES

[1] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cdric Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[2] Vasant Manohar, Stavros Tsakalidis, Pradeep Natarajan, Rohit Prasad, and Prem Natarajan, "Audio-visual fusion using bayesian model combination for web video retrieval," in *ACM Multimedia*, 2011, pp. 1537–1540.

[3] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, pp. 91–110, 2004.

[4] Ivan Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[5] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1980, vol. 28, pp. 357–66.

[6] S. V. N. Vishwanathan, Z. Sun, N. Theera-Ampornpunt, and M. Varma, "Multiple kernel learning and the SMO algorithm," in *Advances in Neural Information Processing Systems*, December 2010.

[7] Guangnan Ye, I-Hong Jhuo, Dong Liu, Yu-Gang Jiang, D.T. Lee, and Shih Fu Chang, "Joint audio-visual bimodal codewords for video event detection," *ACM ICMR*, 2012.

[8] Wei Jiang and Alexander C. Loui, "Audio-visual grouplet: temporal audio-visual interactions for general video concept classification," in *Proceedings of the 19th ACM international conference on Multimedia*, New York, NY, USA, 2011, MM '11, pp. 123–132, ACM.

[9] Wei Jiang, Courtenay Cotton, Shih-Fu Chang, Dan Ellis, and Alexander Loui, "Short-term audio-visual atoms for generic video concept classification," in *Proceedings of the 17th ACM international conference on Multimedia*, New York, NY, USA, 2009, MM '09, pp. 5–14, ACM.

[10] Inderjit S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2001, KDD '01, pp. 269–274, ACM.

[11] J.R. Bellegarda, J.W. Butzberger, Yen-Lu Chow, N.B. Coccaro, and D. Naik, "A novel word clustering algorithm based on latent semantic analysis," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, may 1996, vol. 1, pp. 172 –175 vol. 1.

[12] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Wessel Kraaij, Alan F. Smeaton, and Georges Quenot, "Trecvid 2011 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2011*. NIST, USA, 2011.

[13] Jan C. van Gemert, Cor J. Veenman, Arnold W. M. Smeulders, and Jan-Mark Geusebroek, "Visual word ambiguity," *IEEE PAMI*, vol. 32(7), pp. 1271–1283, 2010.

[14] Jingen Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, june 2011, pp. 3209 –3216.

[15] Pierre Tirilly, Vincent Claveau, and Patrick Gros, "Language modeling for bag-of-visual words image categorization," in *Proceedings of the 2008 international conference on Content-based image and video retrieval*, New York, NY, USA, 2008, CIVR '08, pp. 249–258, ACM.

[16] Jianbo Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888 –905, aug 2000.