UNSUPERVISED ANCHOR SHOT DETECTION USING MULTI-MODAL SPECTRAL CLUSTERING

Chengyuan Ma and Chin-Hui Lee

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
{cyma, chl}@ece.gatech.edu

ABSTRACT

This paper presents a novel unsupervised method for anchor shot detection using spectral clustering with multi-modal features. Unlike previous unsupervised studies where the acoustic trajectory features can not be combined with visual features directly, only a pairwise distance matrix from each attribute is needed instead of individual samples so that diverse information from heterogeneous features can be integrated in a unified manner. Experimental evaluation on a subset of the TRECVID 2004 dataset showed that an appropriate incorporation of the acoustic information with visual information will improve the F1 score from 0.68 for visual information only system to 0.87 in our unsupervised anchor shot detection system. Also a comparison study on the same dataset with a supervised system showed that the performance of our unsupervised system approach that of the supervised system.

Index Terms— anchor shot detection, spectral clustering

1. INTRODUCTION

Broadcast news video is well structured: from frames to shots, scenes and stories. To make the implicit structure explicit is crucial in content-based video indexing and retrieval. Part of the structure information is the change points of many semantic units and events, e.g., video story boundaries and speaker change times. Many studies have been conducted at different levels. By now, the shot boundary detection and keyframe extraction are mostly well-established techniques [1] [2]. In a top-down framework, the broadcast news can be modeled with a hidden Markov model (HMM) [3] or a probabilistic context-free grammar (PCFG) such that the change points or more complicated structure information can be obtained from a decoding procedure or a parsing tree. While in a bottom-up detection-based framework, low-level semantic event and concept detectors are used to reveal the high-level hidden structure. And many discriminative models have been investigated, e.g., support vector machine (SVM) [1] and maximum entropy (ME) model [4] . Semantic concept detection at the shot level is of great importance for high-level video applications, e.g., anchor shot detection for story segmentation, caption and text shot detection for video story summarization, etc. In an anchor shot detection, we try to find all the shots with one or two anchorpersons in a studio background. Our preliminary experiments on a 17-hour broadcast new video dataset show that if the anchor shot location was treated as the news story boundary, this feature alone can achieve an accuracy of 61.7% in story segmentation. Similar results have been verified by many studies. Research work on significant feature selection for story segmentation using an information gain criterion also shows that anchor shot is the most important feature in story segmentation [3].

Previous studies on anchor shot detection include both supervised and unsupervised methods [5][6]. Almost all systems use only visual features and face information from keyframes. To utilize the spatial information in anchor shots, some researchers employed pattern matching with several pre-defined spatial structures [7]. A supervised system using SVM gave an average accuracy of 91.3% on TRECVID 2005 dataset [6]. While, unsupervised anchor shot detection usually result in an average accuracy about 60%-80% [3]. Although supervised systems have better performance for specified channels and programs, and can be used in on-line processing, its limitation is obvious: the production rules and the style are varying from channel to channel and from time to time.

Generally speaking, unsupervised anchor shot detection systems perform clustering on the shots with detected faces using visual features, such as a color histogram [3]. In another unsupervised system, graph-theoretical clustering (GTC) with minimum spanning tree is used with only visual features and a face detector [5]. In all these unsupervised systems, cues from audio track were ignored. Part of the reason is that it is not trivial to find suitable representations for multi-modalities to integrate the heterogeneous features into a unified framework.

In this paper, we propose a novel unsupervised learning framework to represent multi-modal features in a unified and systematic manner. Spectral clustering with multi-modal features from video, audio and high-level information are investigated thoroughly. Experimental results show that fusion of acoustic features indeed provide complementary information and the anchor shot detection performance can approach those results only achievable in supervised systems.

2. SPECTRAL CLUSTERING

Spectral clustering was originated from graph partitioning based on spectral graph theory [8] and has been intensively studied in machine learning community [9].

2.1. Spectral clustering algorithm

Given n vectors, $X = (x_1, x_2, \cdots, x_n), x_i \in \mathbb{R}^d$, a weighted undirected graph $G = (V, E)$ is constructed to encode the neighborhood structure of X. Here V is the vertex set and E is the edge set. Each edge $e_{i,j}$ connecting nodes i and j is associated with a weight $d(i, j) > 0$. An affinity matrix A, is formed for G to represent the pairwise similarity. In practice, the affinity matrix is often obtained

using kernel tricks to project the data into a high dimensional feature space and a Gaussian kernel is the most used one.

$$A_{ij} = \exp(-\frac{d(i,j)}{\sigma^2}) \quad (1)$$

σ is the size of the Gaussian kernel and used as a scaling parameter.

There are many variants of spectral clustering algorithms. The major difference is in the construction of the affinity matrix. Here we follow Ng's algorithm [9].

1. Define $D = \mathrm{diag}(d_1, d_2, \cdots, d_n)$ to be the degree matrix of A, here $d_i = \sum_{j=1}^{n} A_{ij}$. And construct a normalized affinity matrix L as follow,

$$L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \quad (2)$$

2. Find the k largest eigenvectors of L, and form a matrix U by stacking the eigenvectors in columns, $U = [u_1, u_2, \cdots, u_k] \in \mathbb{R}^{n*k}$.
3. Form a matrix R from U by normalizing each row of U to have unit length. A row vector in R is a new feature vector associated with each node. Now All the nodes are on a unit sphere in the spectral space spanned by the k largest eigenvectors.
4. Cluster the rows of R into k clusters with a k-means algorithm or any other clustering algorithms.

2.2. Advantages of spectral clustering

Spectral clustering has many advantages over conventional clustering algorithms. First, kernel techniques are used to project the data into a high-dimensional feature space in which the clusters can be more spatially distinct and compact. Second, spectral clustering is also a nonlinear dimensionality reduction method. The actual clustering process is performed in a low-dimensional spectral space spanned by the first k largest eigenvectors of the normalized affinity matrix. It's more efficient and robust for initialization. Third, theoretical analysis shows that spectral decomposition can reveal the block structure of the affinity matrix, which is related to the number of intrinsic clusters [9]. Finally, in spectral clustering, only a distance matrix is needed instead of individual samples for each attribute or a centroid of a cluster. For instance, it's hard to find a centroid of a group of audio segments with different durations. So it's difficult to integrate acoustic features into conventional k-means based clustering algorithms. Whereas, the shot-wise distance matrix from acoustic features or other attributes can be obtained easily. These properties make spectral clustering a better choice for integrating heterogeneous information encoded in attribute distance matrices in unsupervised anchor shot detection.

3. MULTI-MODAL FEATURE REPRESENTATION

For anchor shot detection, there is rich of information about anchors embedded in the video and audio signals. Next we describe the visual, acoustic and high-level features used in this paper.

3.1. Visual feature representation

Intuitively, visual similarity is the most salient cue for anchor shot detection. Also the spatial structure of the keyframes from anchor shots are helpful. The following visual features have been shown to be effective in TRECVID evaluations [10] and the exact feature extraction procedures are adopted in this paper.

Color histogram Color histogram represent the global color distribution and it is invariant to affine transformation.

Grid color moment Grid color moment is used to compensate the lackness of local information in color histogram. By this way, part of the spatial information was encoded into visual feature vectors.

Edge direction histogram An edge direction histogram represents the distribution of directions of edges in an image.

Gabor filter texture Gabor filter banks are used for texture feature analyzing and representation.

3.2. Audio feature representation

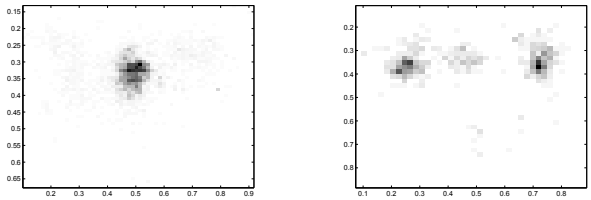
A Gaussian mixture model (GMM) was estimated from audio track for each shot segment. Let O_i be the acoustic observations associated with the ith shot segment, and M_i be the acoustic model estimated from O_i. The normalized cross log-likelihood ratio (CLLR) [11] is defined as following,

$$\mathrm{CLLR}(i,j) = \left| \frac{1}{N_i} \log \frac{p(O_i|M_i)}{p(O_i|M_j)} + \frac{1}{N_j} \log \frac{p(O_j|M_j)}{p(O_j|M_i)} \right| \quad (3)$$

The CLLR is a symmetric and non-negative dissimilarity measure between O_i and O_j. It's has been used in speaker identification and speaker segmentation.

3.3. High-level feature representation

Face detection is conducted by a cascade of a collection of weak classifiers [12]. For one-anchor shots, a Gaussian model M_{one} is estimated to represent the position of the detected face regions and face sizes. While for two-anchor shots, a GMM M_{two} was constructed in a similar way. Figure 1 shows the 2-D histograms of detected face positions in one- and two-anchor shots. In our study, these two models are used in target cluster selection and false alarm pruning. Although the face detection result is not perfect with some missing and falsely detected faces. Our experiment results show that 97.1% of the manually labeled anchor shots are covered by the shots with detected faces. On the other hand, the shots with detected faces account for about 30% of all the shots. So working only with shots with detected faces will greatly reduce the computation complexity.



(a) One face (b) Two faces

Fig. 1. 2-D histogram of face position in anchor shots.

4. PREPROCESSING AND POSTPROCESSING

Although spectral clustering is straightforward for simple tasks, when dealing with heterogeneous distance matrices from diverse attributes in anchor shot detection, there are several preprocessing and postprocessing steps needed to be investigated carefully.

4.1. Matrix normalization and combination

The original spectral clustering works on single affinity matrix. We extend it to integrate multiple heterogeneous features. A weighted sum of all normalized affinity (or distance) matrices was used. Both distance and affinity matrix combination schemes were studied. When a Gaussian kernel is used, an addition of normalized distance matrices is equivalent to a multiplication of affinity matrix with different scaling factors:

$$A = \sum_{i=1}^{I} w_i * A_i \quad (4)$$

Where A_i and w_i are the affinity (or distance) matrix and weight for the ith attribute. These weights are critical factors in combination. They reflect the relative importance of each attribute. And they are application dependent and prior knowledge is needed for an appropriate choice. So each w_i is chosen by cross-validation.

In order to make effective feature integration, it's necessary to perform matrix normalization before a matrix combination. To ensure that normalization doesn't change the eigenvectors and the spectral space, a variance scaling normalization is conducted on either distance or affinity matrices. Such normalization procedures make distance (or affinity) matrices from different attributes having the same standard deviation and comparable for easy combination.

4.2. Cluster number selection

Cluster number selection or more general model order selection is a difficult problem. One advantage of spectral clustering is that the eigen-structure of the normalized affinity matrix can reveal the intrinsic structure of the data. In some ideal cases, the multiplicity of eigenvalue 1 is the cluster number. In more general cases, based on the matrix perturbation theory, the eigengap indicating the stability of the eigen-structure is used to determine the number of clusters [9]. A big drop in eigengap could indicate the true cluster number.

$$\delta_k = 1 - \frac{\lambda_{k+1}}{\lambda_k} \quad (5)$$

Where, λ_k is the kth largest eigenvalue and δ_k is the kth eigengap. Also, some other cluster validity scores, such as Rand index and Hubert index can be used for cluster number selection [13]. In our experiment, the cluster number selected by eigengap is not as good as the cross-validation method in terms of detection performance. The cross-validation chooses the cluster number to 3 and 4.

4.3. Target cluster selection

The face information (positions and sizes) x from a face detector and the two estimated face information models M_{one} and M_{two} were used to select the target cluster. For each shot with detected faces, a log-likelihood $p(x|M_{one})$ or $p(x|M_{two})$ is computed depending on the number of detected faces. The standard deviation of the log-likelihoods within each cluster is used as the criterion. The cluster with the minimum standard deviation was selected as the target anchor shot cluster.

4.4. Pruning in target cluster

Within the target cluster, the outlier shots that have extreme log-likelihoods will be removed from the target cluster. Previous studies show that anchor shots usually have a duration longer than 2 seconds. So the duration and motion quantity of each shot were also used to eliminate false alarms to improve the precision.

5. EXPERIMENT SETUP AND RESULT ANALYSIS

A subset from TRECVID 2004 [14] is used for the anchor shot detection evaluation. It consists of 34 video clips with a total length of about 17 hours. The data are CNN and ABC broadcast news video of year 1998. For a reliable evaluation, all the anchor shots were manually labeled for this dataset.

The shot segmentation and keyframe extraction are conducted with a publicly available tool VideoAnnEx [2]. The audio track is demultiplexed from the MPEG stream with 16 KHz sampling rate and 16 bits. Mel frequency cepstral coefficient (MFCC) features were extracted from audio signals in a conventional way [15].

The performance of an anchor shot detection system is usually evaluated with precision and recall measures used in information retrieval. Meanwhile, a single F1 score which combine the recall and precision is also used for performance comparison.

5.1. Comparison of combination strategies

The first experiment was to compare the distance and affinity matrix combination schemes mentioned in section 4.1. Table 1 shows the average performance measure in both schemes. Although the combination of distance matrices had a slightly better F1 score, both Wilcox matched-pairs signed-rank test (p-value $= 0.24$) and the student-t test for one sample (p-value $= 0.19$) can not reject the null hypothesis (the median or mean difference in F1 score between two combination schemes is zero). So we conclude that the difference between the two combination strategies in terms of F1 score is not statistically significant.

Table 1. Two combination strategies.

	Precision	Recall	F1
affinity combination	83.1%	89.7%	0.863
distance combination	86.5%	87.6%	0.871

5.2. Weighting parameters interpretation

The second experiment was to investigate the effects of weighting parameters mentioned in section 4.1 on detection performance. Figure 2 shows the F1 scores under different weighting parameters. Here, μ is a weight for acoustic affinity matrix and $1 - \mu$ is a weight for visual affinity matrix. This value indicates the relative importance of visual features and acoustic features in unsupervised anchor shot detection. When $\mu = 0.0$, it means only visual features were used in clustering, the F1 is about 0.68. When $\mu = 1.0$, it means that only the acoustic feature was used in clustering, the F1 is about 0.52. While when $\mu = 0.35$, F1 peaks at about 0.87. We can draw several conclusions from this figure. First, the visual feature is more significant than acoustic feature for anchor shot detection. Second, when visual features are combined with acoustic features with an appropriate weight, the detection performance can be improved dramatically. It demonstrates the effectiveness of combining of heterogeneous features.

5.3. Effectiveness of pruning

The third experiment was to verify the effectiveness of the false alarm pruning strategies mentioned in section 4.4 and the experiment result is listed in Table 2. Similar hypothesis testing procedures showed that the F1 score after false alarm pruning is significantly

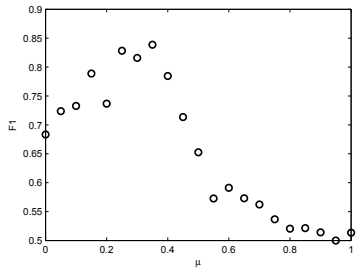


Fig. 2. F1 score vs. weighting factor μ.

better than that before pruning. It's clear that pruning can greatly improve the precision from 79.9% to 86.5%, while the recall will be slightly reduced from 89.4% to 87.6%.

Table 2. Effectiveness of pruning.

	Precision	Recall	F1
w/o pruning	79.9%	89.4%	0.844
w/ pruning	86.5%	87.6%	0.871

5.4. Comparison with a supervised system

The fourth experiment was to compare the proposed unsupervised anchor detection system with a supervised system using SVM [6] and the LIBSVM [16] toolkit was used in our implementation. Table 3 shows the performance of our supervised system. It's clear that the performance of our unsupervised system (F1 = 0.871) approach the performance of the supervised system (F1 = 0.891). Also we have tried to incorporate the acoustic information into the supervised system to further improve the performance. The acoustic information of each shot can not be fused with visual features directly in supervised system. Because we are building a general anchor shot model, not for a specified anchorperson, for each shot, the statistics from the 20 smallest acoustic distance were computed and used as acoustic features for this shot. Experiment results in Table 3 show that the integration of shot-wise acoustic distance can further improve the performance of the supervised system. Although the improvement in F1 score is not significant, the precision has been significantly improved. Some of the reason is that by incorporating the statistics from acoustic distances, part of the global consistency for anchor shots within each video clip was captured instead of only local visual features from single shot.

Table 3. Supervised anchor shot detection.

	Precision	Recall	F1
visual	93.3%	85.3%	0.891
visual + acoustic	95.1%	85.8%	0.902

6. SUMMARY

In this paper, we proposed a novel unsupervised framework for anchor shot detection in broadcast news video. By using a spectral clustering algorithm in which only the pairwise distance matrix for each attribute is needed, diverse information from heterogeneous features can be integrated in a unified manner. Experiment results show that such an integration can greatly improve the performance of the unsupervised system. And the performance of the unsupervised system can approach the performance of the supervised system.

7. ACKNOWLEDGEMENT

The authors are grateful to Shi Yong Neo at Nation University of Singapore for sharing the evaluation dataset and to Mamadou Diao at Georgia Tech for providing part of the ground-truth labels for this dataset.

8. REFERENCES

[1] W. Hsu, L. S. Kennedy, S.-F. Chang, M. Franz, and J. R. Smith, "Columbia-IBM news video story segmentation in TRECVID 2004," Tech. Rep., Columbia University, 2005.

[2] C.-Y. Lin, B. L. Tseng, and J. R. Smith, "VideoAnnEx: IBM MPEG-7 annotation tool," in *Proc. of ICME*, 2003.

[3] T.-S. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu, "Story boundary detection in large broadcast news video archives: techniques, experience and trends," in *Proc. of ACMMM*, 2004, pp. 656–659.

[4] W. Hsu and S.-F. Chang, "A statistical framework for fusing mid-level perceptual features in news story segmentation," in *Proc. of ICME*, 2003.

[5] M. De Santo, P. Foggia, G. Percannella, C. Sansone, and M. Vento, "An unsupervised algorithm for anchor shot detection," in *Proc. of ICPR*, 2006.

[6] A. Yanagawa, W. Hsu, and S.-F. Chang, "Anchor shot detection in TRECVID-2005 broadcast news videos," Tech. Rep., Columbia University, 2005.

[7] H. Zhang, Y. Gong, S. W. Smoliar, and S. Y. Tan, "Automatic parsing of news video," in *Proc. of ICMCS*, 1994.

[8] F. R. K. Chung, *Spectral Graph Theory*, Amer. Math. Society, 1997.

[9] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, 2001.

[10] A. Yanagawa, W. Hsu, and S.-F. Chang, "Brief descriptions of visual features for baseline TRECVID," Tech. Rep., Columbia University, 2006.

[11] C. Ma, P. Nguyen, and M. Mahajan, "Finding speaker identities with a conditional maximum entropy model," in *Proc. of ICASSP*, 2007, pp. 261–264.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of CVPR*, 2001.

[13] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.

[14] W. Kraaij, A. F. Smeaton, P. Over, and J. Arlandis, "TRECVID 2004 - an overview," Tech. Rep., National Institute of Standards and Technology, 2005.

[15] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognition and Artificial Intelligence*, pp. 374–388, 1976.

[16] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001.