JOINT INTER AND INTRA SHOT MODELING FOR SPECTRAL VIDEO SHOT CLUSTERING

Jianning Zhang, Lifeng Sun, Shiqiang Yang, Yuzhuo Zhong

Department of Computer Science and Technology, Tsinghua University, China

ABSTRACT

This paper proposed a novel video shot clustering algorithm using spectral method by joint modeling of inter and intra shot. Gauss Mixture Model (GMM) is used for probabilistic space-time modeling of intra-shot pixels. The spectral clustering method is applied on the GMM parameters. The problem of automatic model selection is currently an open issue for spectral method. Here we propose a novel automatic model selection based on joint inter-intra model optimization to achieve the global optimization in both model parameters and cluster numbers. We compare the proposed method with the conventional spectral method for sports video clustering. The simulation results show more accuracy on how many clusters and clustering results.

1. INTRODUCTION

Segmentation and grouping of video shots are the important basic technologies in content-based video analysis [1,2,3]. Video shots clustering technologies allow unsupervised video content learning and organizing without any labeling by hand. So the clustering methods are deeply studied in recent years and widely applied on video content analysis and retrieval.

The video shot is the minimal semantic unit, which has coherent chrominance and continuous objects moving without camera switching. The video frames in one shot can be considered to be similar appearance and have spatial-temporal relations. So the content of each shot is the basic meaningful semantic unit for further analysis. The grouping of video shots can achieve the classes with the similar represents in both low-level features and semantics. There are two mainly categories of grouping methods: supervised methods and unsupervised methods.

One popular kind of video shots grouping methods uses supervised learning [4,5]. The low-level features are extracted from each shot and the training data is labeled by hand. The neural network and SVM can be used for training the classifiers. The probabilistic models are also used for training and classifying, such as Hidden Markov Model in [5]. The supervised methods are more accurate and efficient than unsupervised methods. But the learning needs much time spending work of labeling. The learned classifiers or models can only be applied on the same sets of videos. So different classifiers or models should be trained for different video sets.

To overcome these problems, the clustering methods for unsupervised learning are developed for video shots grouping. The clustering methods can apply directly on the video data without any labeling. And these methods are universal solutions for different sets of video grouping. The simple but efficient clustering algorithm is the K-mean algorithm. In [10], a probabilistic hierarchical clustering using GMM is proposed. The spectral clustering methods [7,8] are popular studied in recent years for image segmentations. It is shown of better clustering results than methods in [6] and [10] for video shots grouping in [9].

The features used for represent each shot are commonly selected using key frame or average color histogram. But these features are not sufficient for representing the objects distribution and dynamics.

The automatic cluster number selection is still an open issue for clustering methods. The mean squared error measures are not efficient analyzed in [9]. And [9] proposes the measures for spectral clustering using eigen-gap and relative cuts. The minimum description length (MDL) measurement is applied for probabilistic models in [10]. These methods only concern the data distribution in feature space and lack the consideration of the semantic representation for video shots. The dimension of feature space can affect the clustering results obviously and is not considered in these measures.

In this paper, we propose a joint method based on spectral clustering. The proposed method using GMM to represent the intra shot features, which can make more description of the objects distribution and dynamics in one shot than key frame or average histogram. The spectral

This work has been supported by the National Natural Science Foundation of China under Grant No. 60273008 and by the National High-Tech Research and Development Plan of China (863) under Grant No. 2002AA119040.

0-7803-9332-5/05/$20.00 ©2005 IEEE

clustering is applied for inter shot grouping. To consider the measures of both cluster number and feature-space dimension, we propose a joint automatic model selection for GMM and spectral model.

The paper is organized as follows. Section 2 described our proposed method in detail. The simulation results on sports videos are presented in Section 3. Section 4 draws conclusions.

2. THE PROPOSED METHOD

The proposed method can be illustrated in three steps: GMM modeling of intra-shot pixels, shot clustering using spectral method and joint automatic model selection. The shot segmentation for video sequences is done before the proposed clustering method using the common method with changing detection of color histogram. Then the clustering process is applied on video shots, as shown below in detail.

2.1 Intra-shot modeling

The video shot consists of successive video frames. The pixel values in these frames have spatial and temporal correlations. The distribution and variation of pixels represent the objects and their dynamics in video shot. A probabilistic model of the distribution and variation can be the best feature to represent the shot content.

The probabilistic model we used is GMM. The pixels distribution and dynamics are modeled by mixture gausses in both space and time dimensions. So each pixel is represented by a 6-dimension vector (r,g,b,x,y,t). (r,g,b) is the pixel value in color space. (x,y) is the spatial position and t is the time of the video frame which the pixel lies in. It is shown that the GMM can model the pixel values in both space and time. The maximum likelihood estimation using EM algorithm can be processed to meet the least squared error measurement.

Set the GMM parameter set $\theta = \{\alpha_j, \mu_j, \Sigma_j\}_{j=1}^{k}$, which consists of weight, mean, and covariance of each gauss component. X denotes the input vector of each pixel in one shot. Set $X = (r, g, b, x, y, t)$ and the density function of the mixture k Gaussians is:

$$f(X \mid \theta) = \sum_{j=1}^{k} \alpha_j \frac{1}{\sqrt{(2\pi)^d |\Sigma_j|}} \exp\{-\frac{1}{2}(X-\mu_j)^T \Sigma_j^{-1} (X-\mu_j)\} \quad , (1)$$

where d is the dimension of vector X. If n denotes the number of pixels in one shot, the maximum likelihood estimation of θ is:

$$\hat{\theta} = \arg\max_{\theta} \sum_{j=1}^{n} \log f(X_j \mid \theta) \qquad (2)$$

EM algorithm is applied to estimate the parameter set. The details and its fast algorithm are described in [11].

For each shot, the minimum description length of GMM is calculated to be the model measurement:

$$MDL = \log L(\hat{\theta} \mid X_1 ... X_n) - \frac{l_k}{2} \log n \quad , (3)$$

where L() denotes the likelihood of the input vectors, l_k is the number of parameters needed for the model with *k* Gaussians, given by:

$$l_k = (k-1) + kd + k(\frac{d(d+1)}{2}) \qquad (4)$$

The DL measurement considers both the likelihood of input vectors and the description length of model parameters. It will be used in joint model selection step to select suitable gauss numbers.

2.2 Spectral inter-shot clustering

The spectral method uses affinity matrix to model the similarity of video shots, and clustering on the matrix using eigen-vectors.

The GMM parameters are used to be features extracted for each shot. The distance measure between two shots is defined as:

$$d_{ij} = \sum_{c=1}^{k} |\alpha_{ic}\mu_{ic} - \alpha_{jc}\mu_{jc}| + \sum_{c=1}^{k} |\alpha_{ic}\Sigma_{ic} - \alpha_{jc}\Sigma_{jc}| \quad , (5)$$

where Σ is listing as a vector and distance of vectors is calculated. For the GMM-pair i and j, the gauss component matching and resorting must be done before calculating the distance. The gauss component matching process continuously selects the most similar gauss component pair from these two GMMs with resorting in the same order. So the affinity matrix of all shots is derived from distances. For all i and j shots:

$$A_{ij} = \exp^{-d_{ij}^2 / 2\sigma^2} (i \neq j), A_{ii} = 0 \quad , (6)$$

where σ is a scale parameters. The spectral algorithm is simply described below:

1) Define D(A) to be the diagonal matrix derived from A, that is $D(A)_{ii} = \sum_j A_{ij}$. Then construct L(A) by

$$L(A) = (D(A))^{-1/2} A (D(A))^{-1/2} . \qquad (7)$$

2) Find the *c* largest eigenvectors of L(A). and form the matrix C by stacking the eigenvectors in column.

3) Form the matrix Y from C by normalizing each row to have unit length. The row Y_i is the new feature associated with shot i.

4) Cluster the row Y_i into *c* clusters using K-means.

5) Assign to each shot i the cluster number corresponding to its row.

The cluster number *c* should be decided as shown in next subsection. We can see that the *c* largest eigenvectors denote *c* orthogonal directions in eigenspace, so the suitable *c* will lead the best departments of data in eigenspace.

2.3 Joint model selection

The model selection includes automatic selection of number *k* and number *c*. *k* is the number of Gaussians for intra-shot, and *c* is the number of clusters derived from spectral algorithm.

In order to select the best suitable *k*, minimum description length for GMM of shot i can be derived by (3) and the best *k* will be selected for maximizing the average MDL values for all shots.

$$k = \arg\max_{k} \left(\frac{\sum_{i=1}^{n} MDL_{ik}}{n} \right) \qquad (8)$$

So the measurement of intra-shot model can be defined as:

$$M_{\mathrm{int}\,ra}(k) = \frac{\sum_{i=1}^{n} MDL_{ik}}{n} \qquad (9)$$

In [9], the eigengap and the relative cut are used to measure the performance of spectral models. And best cluster number *c* is selected using separate thresholds.

The eigengap is used to measure the stability of a matrix. So the eigengap shows the stability of each cluster. The eigengap is defined as:

$$M_{gap}(c) = \min_{m \in 1 \ldots c} \left(1 - \frac{\lambda_2(m)}{\lambda_1(m)}\right) \quad , (10)$$

where $\lambda_1(c)$ and $\lambda_2(c)$ are the two largest eigenvalues of matrix $L(A_m^{ii})$ extracted from matrix A by selection the shots in cluster m. The smallest c is selected for M_{gap} exceed a threshold.

The relative cut is defined to measure the department of clusters as:

$$M_{rcut}(c) = \frac{\sum_{m=1}^{c} \sum_{l=1, l \neq m}^{c} \left(\sum_{i \in S(m)} \sum_{j \in S(l)} A_{ij} \right)}{\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}} \quad , (11)$$

where S(m) means the set of shots in cluster m. The largest c is selected for M_{rcut} below a threshold.

But [9] uses two thresholds to measure the eigengap and relative cut separately. We can see that when the cluster number *c* increases, the cluster stability and cluster department will both increase. The best *c* will be at the position of maximum ratio of M_{rcut} and M_{gap}.

We also analyze that the relationship of the intra-shot model measurement and inter-shot measurement. The gauss number *k* represents the feature space dimension. If *k* is too small, the features are not sufficient to represent the video shot, so that the spectral clustering will be insufficient too. If *k* is too large, the features will be confused in feature space that makes the cluster number to be large.

By considering the factors above, we proposed a joint model selection measurement to select suitable *k* and *c* in the same step. The new measurement is defined as:

$$M(c,k) = \frac{M_{\mathrm{int}\,ra}(k) * M_{rcut}(c)}{M_{gap}(c)} \qquad (12)$$

The best *c* and *k* is selected by maximizing the M:

$$c,k = \arg\max_{c,k}(M(c,k)) \qquad (13)$$

3. SIMULATION RESULTS

In the experiments, we compare our proposed method with the conventional spectral clustering in [9]. The test videos belong to the NBA basketball videos.

We test total three clips of videos. Each clip video has more than 300 video shots, and the playing shots and the stop shots are all remained.

We select the average color histogram to be the intra-shot features for the conventional spectral method. The scale parameter σ in (6) is set to be 0.15 for both conventional spectral and proposed methods according to [9]. Figure 1 and 2 show the measurement results of clip 2 for cluster numbers from 6 to 34 in conventional spectral method and the proposed method. Mcost means the measurement result. For conventional spectral method, the Mcost is calculated using the relative cut divided by the eigengap. And M(c,k) in (12) is used as Mcost in the proposed method. From Figure 2, we can see that the spectral clustering measurement curves are some different for different number of Gaussian components and the best numbers of Gaussian components can be shown for maximum joint measurement results.

Videos	Method	Cluster number	Total shots	P	R
Clip1	Conventional	30	353	0.9348	0.9235
	Proposed	28	353	0.9547	0.9490
Clip2	Conventional	32	382	0.9529	0.9372
	Proposed	24	382	0.9712	0.9476
Clip3	Conventional	26	490	0.9449	0.9347
	Proposed	16	490	0.9633	0.9489

Table 1. Comparison results

The best cluster number is auto selected and the ground-truth is generated for each method according to the selected cluster number. The experiments results are shown in Table 1, where P means precision and R means recall.

From the table, we can see that our proposed method is better results than the conventional ones. The cluster number of the proposed method is different from the conventional method, which indicates that the feature space dimension affects the cluster number selection. The best cluster number is selected by maximizing the joint likelihood of intra-inter shots modeling.

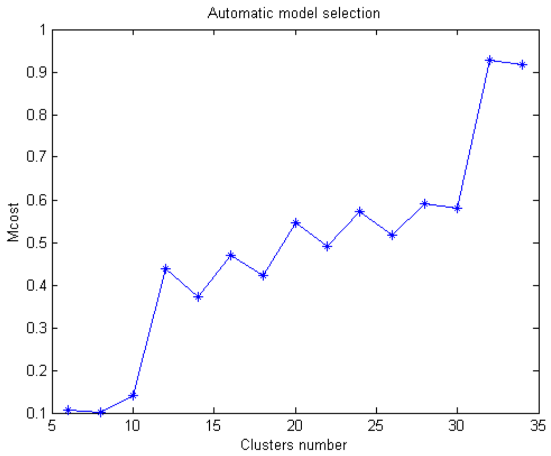

Figure 1. Measurement results for conventional spectral method

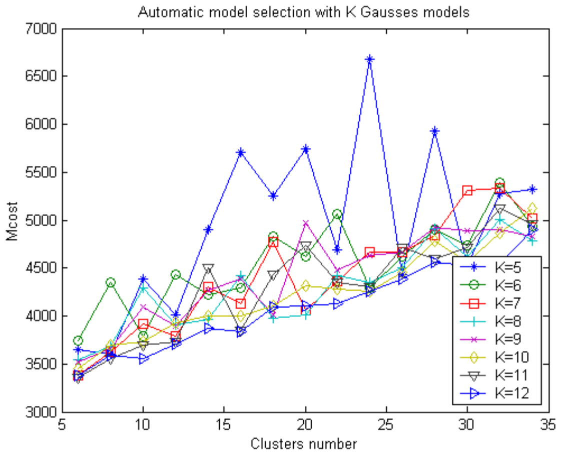

Figure 2. Measurement results for proposed method

4. CONCLUSION

In this paper, we proposed a joint clustering method based on spectral method. The proposed method using GMM to represent the intra shot features, which can make more description of the objects distribution and dynamics in one shot than key frame or average histogram. The spectral clustering is applied for inter shot grouping. To consider the measures of both cluster number and feature-space dimension, we propose a joint automatic model selection for GMM and spectral model. The simulation results show the more clustering accuracy than the conventional spectral algorithm. And the join intra-inter shot model selection can achieve the global optimization for feature selection and clustering number decision. The future work will focus on the intra-shot pattern discovery based on the proposed spectral clustering. The model will be improved for representing the dynamics of intra-shot objects.

11. REFERENCES

[1] D. Gatica-Perez, M.-T. Sun, and A. Loui, "Consumer Video Structuring by Probabilistic Merging of Video Segments," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, Aug. 2001.

[2] Y. Rui and T. Huang, "A Unified Framework for Video Browsing and Retrieval," in *Image and Video Processing Handbook*, Academic Press, pp.705-715, 2000.

[3] M. Yeung, B.L. Yeo, and B. Liu, "Segmentation of Video by Clustering and Graph Analysis," *Computer Vision and Image Understanding*, Vol. 71, No. 1, pp. 94-109, July 1998.

[4] L. Xie, S.-F. Chang, A. Divakaran and H. Sun, "Structure Analysis of Soccer Video with Hidden Markov Models", *Proc. Interational Conference on Acoustic, Speech and Signal Processing*, Orlando, FL, USA, May 13-17, 2002.

[5] E. Kijak, L. Oisel, P. Gros, "Hierarchical structure analysis of sport videos using hmms", *ICIP 2003*, Volume: 2 , Sept. 14-17, 2003, Pages:1025 – 1028

[6] J.-M. Odobez, D. Gatica-Perez and M. Guillemot, "On Spectral Methods and Structuring of Home Videos," *IDIAP Technical Report*, IDIAP-RR-55, Nov. 2002.

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

[8] A. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Proc. NIPS*, Dec 2001.

[9] J.M. Odobez, D. Gatica-Perez, M. Guillemot, "video shot clustering using spectral methods," in *3th Workshop on Content-Based Multimedia Indexing(CBMI)*, sept 2003.

[10] D.G. Perez, A. Loui, M.T. Sun, "Finding Structure in Home Video by Probabilistic Hierarchical Clustering, " *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 13, No. 6, Jun 2003.

[11] H. Greenspan, J. Goldberger, A. Mayer, "Probabilistic Space-Time video Modeling via Piecewise GMM", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 26, No.3, Mar 2004.