Video Shot Segmentation Using Graph-based Dominant-Set Clustering

Li Li
National Laboratory of Pattern Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, China
lli@nlpr.ia.ac.cn

Xianglin Zeng
National Laboratory of Pattern Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, China
xlzeng@nlpr.ia.ac.cn

Xi Li
National Laboratory of Pattern Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, China
lixi@nlpr.ia.ac.cn

Weiming Hu
National Laboratory of Pattern Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, China
wmhu@nlpr.ia.ac.cn

Pengfei Zhu
National Laboratory of Pattern Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, China
pfzhu@nlpr.ia.ac.cn

ABSTRACT

Video shot segmentation is a solid foundation for automatic video content analysis, for most content based video retrieval tasks require accurate segmentation of video boundaries. In recent years, using the tools of data mining and machine learning to detect shot boundaries has become more and more popular. In this paper, we propose an effective video segmentation approach based on a dominant-set clustering algorithm. The algorithm can not only automatically determine the number of video shots, but also obtain accurate shot boundaries with low computation complexity. Experimental results have demonstrated the effectiveness of the proposed shot segmentation approach.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing-indexing methods; I.2.10 [**Artificial Intelligence**]

General Terms

Algorithm, Experimentation

Keywords

Shot boundary detection, dominant-set clustering

1. INTRODUCTION

A large amount of digital videos have been generated due to the rapid development of computing and network infrastructures. How to effectively and efficiently manage numerous video information is crucial for content-based video browsing and retrieval. In general, videos can be represented by a hierarchical structure, while shots are the basis units for constructing high level semantic scenes. Thus, shot boundary detection is an important preprocessing step for efficient browsing and further content analysis. A shot consists of consecutive frames which are usually captured by a single camera action. Typically, there are no significant content changes between successive frames in the shot [6]. Based on the identified shots, high-level scene analysis and video abstraction can be performed to construct a better representation of a video.

Many efforts have been devoted to shot boundary detection, especially, shot boundary detection is a core task at TRECVID[1], and the practice of TRECVID greatly promoted the development of shot boundary detection techniques. Yuan [9] *et al.* categorized the existing shot boundary detection methods into rule-based ones and machine learning based ones. The rule-based methods need some thresholds to detect shot boundaries [10], however, it is difficult to determine proper threshold for various kinds of videos. To avoid the data dependency of threshold selection, the tools of machine learning have been utilized. They are divided into generative and discriminative methods. Since the generative methods highly depend on the assumptions of prior information, the discriminative classifier is preferred. Different discriminative approaches have been applied to shot boundary detection. KNN [3], support vector machines (SVMs) [7] and graph partition model [9] as the supervised approaches also have been used. Cooper [3] built intermediate features via pairwise similarity which incorporate the frame index score, and the intermediate features are classified to detect shot boundary using a binary kNN classifier. Yuan [9] presented a comprehensive review of the existing methods and conducted a unified shot boundary detection

ICIMCS'09 November 23-25, 2009, Kunming, Yunnan, China.

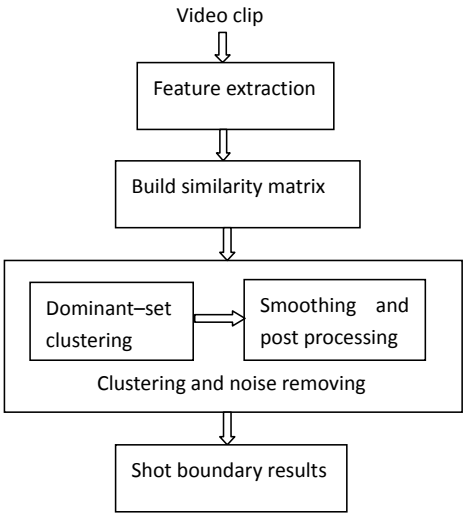

Figure 1: The framework of the proposed shot detection algorithm

system based on a graph partition model.

However, the supervised learning methods share a problem that they rely heavily on a well-chosen training set. Thus, we focus on unsupervised clustering approach to avoid this problem in this paper. K-means[2][5][6] is the most widely used ones due to its simplicity and efficiency, but the requirement of prior knowledge of the number of clusters limits the application of K-means, in particular, the number of shots is not known as far as the problem of shot boundary detection is concerned. Recently, an unsupervised clustering technique called dominant-set clustering [8] attracts significant attention due to its intuitiveness, strong theoretical fundamentals and inherent hierarchical nature. The number of clusters can be dynamically decided without any prior knowledge, thus it is fit to conduct shot boundary detection. One cluster is associated with a shot. The frames of the same cluster exhibit higher similarities while those of different clusters show lower ones. We are pleasured with dominant-set clustering because it considers internal homogeneity and external inhomogeneity simultaneously. Figure 1 illustrates the framework of the proposed method. Based on dominant-set clustering, we propose an effective shot boundary detection approach. Experiments at TRECVID 2005 benchmark demonstrate the effectiveness of the proposed approach.

The remainder of this paper is organized as follows. Section 2 introduce the clustering algorithm based on the concept of dominant set. Shot boundary detection structure based on dominant-set clustering is presented in Section 3. Experimental results are demonstrated in Section 4, followed by conclusions in Section 5.

2. DOMINANT-SET CLUSTERING

2.1 Concept of Dominant Set

Dominant set, defined by Pavan *et al.* [8], is a combinatorial concept in graph theory that generalizes the notion of a maximal complete subgraph to edge-weighted graphs. It simultaneously emphasizes on internal homogeneity and external inhomogeneity, and thus is considered as a general

Input: the similarity matrix $\mathbf{S}$
1. Initialize $\mathbf{S}^k, k = 1$ with $\mathbf{S}$
2. Calculate the local solution of (1) by (2): $\boldsymbol{x}^k$ *and* $f(\boldsymbol{x}^k)$
3. Get the dominant set: $\boldsymbol{D}^k = \sigma(\boldsymbol{x}^k)$
4. Split out $\boldsymbol{S}^k$ from $\mathbf{S}^k$ and get a new similarity affinity matrix $\mathbf{S}^{k+1}$
5. If $\mathbf{S}^{k+1}$ is empty, break, else $\mathbf{S}^k = \mathbf{S}^{k+1}$ and $k = k + 1$, then go to step 2;
Output: $\cup_{l=1}^{k}\{\boldsymbol{D}^l, \boldsymbol{x}^l, f(\boldsymbol{x}^l)\}$

Table 1: Dominant-set clustering algorithm

Input: Affinity vector $\boldsymbol{\alpha} \in \mathbb{R}^n$, $\cup_{l=1}^{k}\{\boldsymbol{D}^l, \boldsymbol{x}^l, f(\boldsymbol{x}^l)\}$
1. $m^l = \frac{\lvert\boldsymbol{D}^l\rvert-1}{\lvert\boldsymbol{D}^l\rvert+1}(\frac{\boldsymbol{\alpha}^T\boldsymbol{x}^l}{f(\boldsymbol{x}^l)} - 1), l \in \{1, \cdots, k\}$
2. Find $l^* = \text{argmax}_l\ m^l$
3. If $m^{l^*} \leq 0$, $l^* = 0$
Output: l^*

Table 2: Dominant-set fast assignment algorithm

definition of "cluster". Pavan *et al.* [8] establish an intriguing connection between the dominant set and a quadratic program as follows:

$$\begin{aligned} max \quad & f(\mathbf{x}) = \mathbf{x}^T\mathbf{S}\mathbf{x} \\ s.t. \quad & \mathbf{x} \in \Delta \end{aligned} \tag{1}$$

where

$$\Delta = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0 \,\text{and}\, \sum_{i=1}^{n} x_i = 1\}$$

where $\mathbf{S}$ is the similarity matrix. Let $\boldsymbol{u}$ denote a strict local solution of the above program. It has been proved by [8] that $\sigma(\boldsymbol{u}) = \{i|u_i > 0\}$ is equivalent to a dominant set of the graph represented by $\mathbf{S}$. In addition, the local maximum $f(\boldsymbol{u})$ indicates the "cohesiveness" of the corresponding cluster. *Replicator equation* can be used to (1):

$$x_i(t+1) = x_i(t)\frac{(\mathbf{S}\mathbf{x}(t))_i}{\mathbf{x}(t)^T\mathbf{S}\mathbf{x}(t)} \tag{2}$$

2.2 Dominant-Set Clustering Algorithm

The concept of dominant set provides an effective framework for iterative pairwise clustering. Considering a set of samples, an undirected edge-weighted graph with no self-loops is built in which each vertex represents a sample and two vertices are linked by an edge whose weight represents the similarity of the two vertices. To cluster the samples into coherent groups, a dominant set of the weighted graph is iteratively found and then removed from the graph until the graph is empty. Table 1 shows the clustering process. Different from traditional clustering algorithms, the dominant-set clustering automatically determines the number of the clusters and has low computational cost.

In order to examine a new sample $\mathbf{x}^{new}$, an out-of sample extension for dominant-set clustering is also made by Pavan *et al.* [8]. The fast assignment algorithm is shown in Table 2, where $\boldsymbol{\alpha}$ is an affinity vector containing the similarities between the new sample $\mathbf{x}^{new}$ and n existing samples. If the output $l^\star > 0$, assign $\mathbf{x}^{new}$ to cluster $l^\star$; else consider $\mathbf{x}^{new}$ as an outlier.

3. VIDEO SHOT SEGMENTATION

As mentioned in Section 1 and Section 2, dominant-set clustering is an unsupervised clustering technique, whose

formal properties and intuitive arguments make it reasonable candidate for shot boundary detection. The architecture of the proposed approach to shot boundary detection is shown in Figure 1. Specifically, the proposed method is composed of four functional blocks: feature extraction, similarity matrix computation, dominant-set clustering and post processing. Each of functional blocks are described in the following sections in detail.

3.1 Feature extraction

Yuan [9] summarized the methods of visual content representation for video shot boundary detection. Several experimental results have shown that the simple histogram metric usually exhibits a satisfactory performance. Since color histogram feature is robust to camera motion as well as object motion, moreover, it gets a better tradeoff between the invariance and the sensitivity of various representation approaches compared to some complicated features (e.g., edge features). Therefore, we employ $16 \times 4 \times 1$ the HSV histogram features, which are invariant to illumination.

3.2 Similarity analysis

The goal of shot boundary detection is to temporally segment the video into some consecutive shots. Not only the content similarity but also the temporal constraint should be taken into consideration. The left of Figure 2 shows an example of original similarity matrix without the temporal constraint. It is evident that frames (i.e., 1 to 126 and 285 to 630) having high similarities are easily grouped into the same shot based on clustering algorithm since weak temporal constraint. Hence, we incorporate the frame index scores into the computation of the similarity matrix with a Gaussian kernel [3]. The similarity between frame i and frame j is computed by:

$$w_{i,j} = \sum_k \min(H_k^i, H_k^j) \times exp(-\frac{1}{d} \cdot \| \frac{i-j}{\sigma} \|^2) \quad (3)$$

where H_k^i is the kth bin color histogram of the ith frame; σ and d are the decay factor which reflect the decreasing rate of the similarity with the temporal interval increasing. The similarity matrix example which is restricted by temporal relation is demonstrated in the right of Figure 2. In this way, the similarity matrix S is built.

3.3 Video shot segmentation

We employ the clustering approach to video shot segmentation with the assumption that the frames in one cluster constitute a shot. Besides, the dominant-set clustering approach is based on the similarity matrix, it is usually easy to cluster the frames which may not be consecutive in frame index into the same shot due to their high similarity, because the clustering approach partly or fully ignores the temporal information. Thus, after conducting the standard dominant-set clustering process, we introduce a smoothing and elimination process:

1. If the cluster label of a frame is not coincides with its two adjacent frames, we treat it as noise with normal smoothing technique such as median filtering.

2. If one cluster contains discontinuous frames sequence, partition it into several segments such that the frames in one shot is temporal contiguous.

4. EXPERIMENTS RESULTS

4.1 Test videos and evaluation criteria

In this paper, we carry out our shot boundary detection experiments on the platform of TRECVID [1] which provides a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results, and shot boundary detection is one of the evaluation tasks. In order to demonstrate the effectiveness of the superiority of dominant-set clustering, we selected seven video clips including soaps, CCTV, CNN, NBC *etc* videos from TRECVID 2005. The resolution of all frames is 352×240. Various difficulties aspect, including camera or object motion, flashing and special editing effects, are within the test video clips.

For evaluation, we employ recall and precision criteria similar to other information task

$$Recall = \frac{\text{The number of shot correctly detected}}{\text{The number of total shots}} \quad (4)$$

$$Precision = \frac{\text{The number of shot correctly detected}}{\text{The number of total shots detected}} \quad (5)$$

In addition, in order to rank performance of different algorithms, F_1 measure is used, F_1 measure combining recall and precision with equal weight in the following form:

$$F_1 = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \quad (6)$$

4.2 Performance

As mentioned in Section 1, the existing methods of shot boundary detection are classified into threshold based and machine learning based. To avoid the problem of choosing proper thresholds and training samples, we chose clustering-based shot detection method for comparison instead of threshold based algorithms. A sequence of consecutive frames are clustered into a shot having high content similarity, while frames from two distinct shots show low similarity, and it is easy to see that only the first and last frames are likely to contain shot boundaries. That is the basic ideas of clustering based shot boundary detection approach. K-means and its variation as the basic clustering have been commonly employed for shot boundary detection because of its simplicity and efficiency. Unlike K-means the dominant-set clustering approach can automatically determine the number of clusters which is shots for the problem of shot detection, while the prior knowledge of the number of clusters is needed for K-means algorithm, and the initialization cluster centers also affect the clustering results. Lu [5] adopted cluster validity analysis as additional information to perform K-means clustering, and a clustering ensemble strategy [2] is proposed to overcome the challenge with the cost of high computation complexity. In order to evaluate the robustness of the dominant-set clustering, we compare the proposed framework with K-means and fuzzy c-means clustering [4] using the same set of videos. The comparison results is in Table 3.

From the Table 3 we can obviously find that our method outperform the other two in terms of accuracy in detecting the right shot boundaries. For some videos the recall and precision which amount to 100% , the algorithm thus performs near perfect. In addition, the extracted histogram metric is simple and just a few iterations is needed

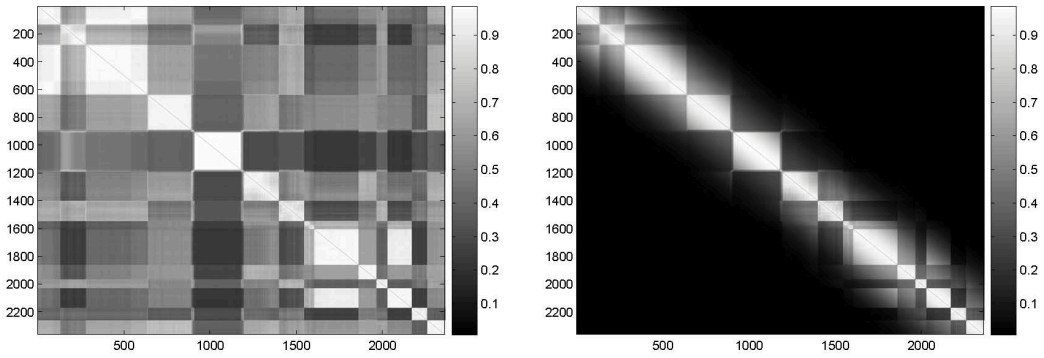

Figure 2: Left:original similarity matrix, right: temporal similarity matrix

	Dominant-set clustering			K-means Clustering			Fuzzy c means clustering		
	recall	precision	F1	recall	precision	F1	recall	precision	F1
CCTV4	100%	91.67%	95.65%	65.45%	75.71%	70.02%	74.55%	62.86%	69.37%
CNN	100%	100%	100%	78.57%	61.11%	68.75%	77.14%	70%	73.40%
MSNBC1	100%	100%	100%	66.67%	86.00%	75.11%	83.33%	65.71%	84.64%
MSNBC2	100%	100%	100%	76.41%	86.92%	81.13%	85.56%	86.62%	86.00%
NBC	82.35%	82.35%	82.35%	77.78%	71.18%	74.33%	75.29%	69.36%	72.41%
Soap1	100%	93.75%	96.77%	60%	77.5%	67.64%	73.33%	80.30%	76.34%
Soap2	100%	100 %	100%	80%	66.67%	72.73%	100%	100%	100%

Table 3: Comparison our propose approach with K-means and Fuzzy c means

for the dominant-set clustering. The high performance of our method comes from two aspect, first and the most important aspect is that the number of clusters or shots can be automatically determined which is a significant consideration for shot detection as well as the number of shots is almost accurate; on the other hand, we integrated the temporal constraint into the shot boundary detection while K-means and fuzzy c means not, therefore, the latter results in false positives. In the experiments, we set $\delta = 100 \sim 200, d = 100$ in most cases for the adaptive clustering because that the video clips contain thousands of frames.

5. CONCLUSION

In this paper, we introduce a novel shot boundary detection method based on dominant-set clustering. The proposed approach can dynamically determine the number of shots with low computational cost. Furthermore, we incorporate the temporal constraint both into the similarity matrix and clustering results, in this way, the temporal contiguous is preserved unlike most clustering-based methods. Experimental results show that the proposed framework has better performance than K-means and Fuzzy c means clustering.

6. ACKNOWLEDGMENT

This work is partly supported by NSFC (Grant No. 60825204 and 60672040) and the National 863 High-Tech R&D Program of China (Grant No. 2006AA01Z453).

7. REFERENCES

[1] Http://www-nlpir.nist.gov/projects/t01v/.

[2] Y. Chang, D. Lee, Y. Hong, and J. Archibald. Unsupervised video shot detection using clustering ensemble with a color global scale-invariant feature tansform descriptor. *EURASIP on Image and Video Processing*, 1-10,2008.

[3] M. Cooper, T. Liu, and E. Rieffel. Video segmentation via temporal pattern classification. *IEEE transactions on multimedia*, 9(3):610–618, April 2007.

[4] C.-C. Lo and S.-J. Wang. Video segmenation using a histogram-based fuzzy c-means clustering algorithm. *Porc. IEEE International Fuzzy Systems*, 2:920–923, December 2002.

[5] H. Lu, Y.-P. Tan, X. Xue, and L. Wu. Shot boundary detection using unsupervised clustering and hypothesis testing. *Proceedings of the International Conference on Communications, Circuits and Systems*, 2:932–936, June 2004.

[6] M. R. Naphade, R.Mehrotra, A. M. Ferman, J. Warnick, T. S. Huang, and A. M. Tekalp. A high-performance shot boundary detection algorithm using multiple cues. In *IEEE Inte. Conf. Image Process*, pages 884–887, 1998.

[7] C.-W. Ngo. A robust dissolve detector by support vector machine. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 283–286, 2003.

[8] M. Pavan and M. Pelillo. A new graph-theoretic approach to clustering and segementation. *Proc. CVPR*, pages 762–768, 2003.

[9] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang. A formal study of shot boundary detection. *IEEE Transaction on circuits and systems for video technology*, 17(2):168–186, February 2007.

[10] H. Zhang, A. Kankanhalli, and S. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1):10–28, 1993.