2008

Key-frame extraction using dominant-set clustering

Xianglin Zeng
Chinese Academy of Sciences, xlzeng@nlpr.ia.ac.cn

Weiming Hu
Chinese Academy of Sciences, wmhu@nlpr.ia.ac.cn

Wanqing Li
University of Wollongong, wanqing@uow.edu.au

Xiaoqing Zhang
Chinese Academy of Sciences, xqzhang@nlpr.ia.ac.cn

Bo Xu
Chinese Academy of Sciences

Publication Details

Zeng, X., Hu, W., Li, W., Zhang, X. & Xu, B. (2008). Key-frame extraction using dominant-set clustering. IEEE International Conference on Multimedia and Expo (pp. 1285-1288). Hannover, Germany: IEEE.

Key-frame extraction using dominant-set clustering

Abstract

Key frames play an important role in video abstraction. Clustering is a popular approach for key-frame extraction. In this paper, we propose a novel method for key-frame extraction based on dominant-set clustering. Compared with the existing clustering-based methods, the proposed method dynamically decides the number of key frames depending on the complexity of video shots, produces key frames in a progressive manner and requires less computation. Experimental results on different types of video shots have verified the effectiveness of the method.

Disciplines

Physical Sciences and Mathematics

Publication Details

Zeng, X., Hu, W., Li, W., Zhang, X. & Xu, B. (2008). Key-frame extraction using dominant-set clustering. IEEE International Conference on Multimedia and Expo (pp. 1285-1288). Hannover, Germany: IEEE.

KEY-FRAME EXTRACTION USING DOMINANT-SET CLUSTERING

Xianglin Zeng, WeimingHu, Wanqing Li†, Xiaoqin Zhang, Bo Xu

National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences
†SCSSE, University of Wollongong, Australia
Email: {xlzeng,wmhu,xqzhang}@nlpr.ia.ac.cn and wanqing@uow.edu.au

ABSTRACT

Key frames play an important role in video abstraction. Clustering is a popular approach for key-frame extraction. In this paper, we propose a novel method for key-frame extraction based on dominant-set clustering. Compared with the existing clustering-based methods, the proposed method dynamically decides the number of key frames depending on the complexity of video shots, produces key frames in a progressive manner and requires less computation. Experimental results on different types of video shots have verified the effectiveness of the method.

Index Terms— Key frames, dominant-set clustering

1. INTRODUCTION

Key frames are a set of salient images extracted from video sequences [1]. They provide a simple yet effective way of summarizing the content of videos for browsing and retrieval and are also widely used in video abstraction due to their compactness. Much research has been conducted in the past few years in understanding the problem of key-frame extraction and developing effective algorithms [2–5]. Truong *et al.* [1] provide a comprehensive overview of the fundamental aspects of the existing approaches and conclude that clustering is one of the effective approaches for key-frame extraction. The clustering approach tends to produce a compact set of key frames and can be performed at both shot level and clip level. In this study, we focus on shot-based key-frame extraction. We adopt the same assumption as in [2] that if the frame is important or salient, the camera will focus more on the scene of the frame. Based on this assumption, a novel method for key-frame extraction is proposed by employing dominant-set clustering algorithm. The efficiency and effectiveness are verified by the experiments on a large set of real videos.

The rest of the paper is organized as follows. Section 2 briefly reviews the related work. Section 3 introduces the clustering algorithm based on the concept of dominant set. Our method for extracting key frames is described in Section 4. Experimental results on real video data are presented and discussed in Section 5, followed by conclusions and remarks in Section 6.

2. RELATED WORK

In clustering-based key-frame extraction, video frames are first grouped into a finite set of clusters in a selected feature space. The selected features are assumed to be able to capture the salient visual content conveyed by the video and the frames closest to the cluster centers are chosen as the key frames, often one frame per cluster. Many traditional clustering algorithms have been explored in the past [2–5]. Zhuang *et al.* [2] employ a sequential clustering technique that assigns the current frame to an existing cluster if their similarity is maximum and exceeds a threshold, or creates a new cluster otherwise. Girgensohn and Boreczky in [3] use the complete link method of hierarchical agglomerative clustering in color feature space. However, both of them are heavily threshold-dependent. Yu *et al.* [4] use fuzzy c-means clustering in the color feature subspace. Gibson *et al.* [5] use Gaussian Mixture Models(GMM) in the eigenspace of the image, in which the number of GMM components is the number of required clusters. The main drawback of the aforementioned methods is that they are not able to automatically determine the number of clusters and, hence, would fail to automatically adapt the clustering to the video content.

Recently, pairwise data clustering techniques, especially the dominant-set clustering, are gaining increasing popularity over traditional clustering techniques due to their intuitiveness, strong theoretical fundamentals and inherent hierarchical nature [6]. This paper employs the dominant-set clustering to extract key frames and the results are compared with the traditional adaptive clustering method in [2].

3. DOMINANT-SET CLUSTERING ALGORITHM

3.1. Concept of Dominant Set

Dominant set, defined by Pavan *et al.* [7], 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 definition of "cluster". Pavan *et al.* [7] establish an intriguing connection between the dominant set and a quadratic program as follows:

978-1-4244-2571-6/08/$25.00 ©2008 IEEE

Table 1. Dominant-set clustering algorithm

```
Input: the similarity matrix $\mathbf{W}$
  1. Initialize $\mathbf{W}^k, k = 1$ with $\mathbf{W}$
  2. Calculate the local solution of (1) by (2): $\boldsymbol{u}^k \; and \; f(\boldsymbol{u}^k)$
  3. Get the dominant set: $\boldsymbol{S}^k = \sigma(\boldsymbol{u}^k)$
  4. Split out $\boldsymbol{S}^k$ from $\mathbf{W}^k$ and get a new affinity matrix $\mathbf{W}^{k+1}$
  5. If $\mathbf{W}^{k+1}$ is not empty, $\mathbf{W}^k = \mathbf{W}^{k+1}$ and $k = k + 1$, then
     go to step 2; else exit
Output: $\cup_{l=1}^{k}\{\boldsymbol{S}^l, \boldsymbol{u}^l, f(\boldsymbol{u}^l)\}$
```

Table 2. Dominant-set fast assignment algorithm

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

$$\begin{array}{ll} max & f(\mathbf{x}) = \mathbf{x}^T \mathbf{W} \mathbf{x} \\ s.t. & \mathbf{x} \in \Delta \end{array} \tag{1}$$

where

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

and $\mathbf{W}$ is the similarity matrix. Let $\boldsymbol{u}$ denote a strict local solution of the above program. It has been proved by [7] that $\sigma(\boldsymbol{u}) = \{i | u_i > 0\}$ is equivalent to a dominant set of the graph represented by $\mathbf{W}$. In addition, the local maximum $f(\boldsymbol{u})$ indicates the "cohesiveness" of the corresponding cluster. *Replicator equation* can be used to solve the program (1):

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

3.2. 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.

To group any new samples after the clustering process has taken place, Table 2 shows the fast assignment algorithm proposed by Pavan *et al.* [6], 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.

4. KEY-FRAME EXTRACTION USING DOMINANT-SET CLUSTERING

Figure 1 shows an overview of the proposed method for key-frame extraction. The method consists of four steps: computing the similarity matrix, clustering, selecting key clusters and selecting key frames.

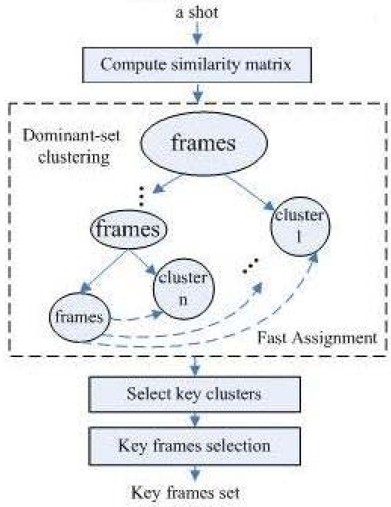


Fig. 1. the flowchart of our approach for key-frame extraction

4.1. Computing the Similarity Matrix

Given a shot, the visual similarity between every pair of frames is calculated in a selected feature space and stored in a matrix. Ideally, the semantic content should be well described in the feature space in order to extract meaningful key frames. However, such a feature space hardly exists. As usual, we calculate the similarity based on low-level visual features such as color, texture and shape. Among different types of low-level visual features, color has been widely used in key-frame extraction due to its robustness and efficiency. Huang *et al.* [8] propose a novel color feature descriptor called autocorrelogram that includes the spatial correlation of colors. They also develop an efficient way to compute the autocorrelogram and propose a difference measure called d_1 distance measure to compare two feature vectors. Previous research has shown that the autocorrelogram and the d_1 distance measure are effective in capturing the visual saliency of video frames. In this paper, we adopt the autocorrelogram in HSV color space as the low-level visual feature. In order to reduce the influence of the illumination variations, colors are quantized into $16 \times 4 \times 1$ levels. We choose the same distance set $D = \{1, 3, 5, 7\}$ as [8]. The d_1 distance measure is used to compute the difference $d_{i,j}$ between frame i and frame j. Then the similarity between frame i and frame j is calculated by (3)

$$w_{i,j} = exp\,(\frac{-d_{i,j}^2}{\delta}) \tag{3}$$

where δ is a positive real number which affects the decreasing rate of w. Notice that other features such as shape and motion information and different distance metrics can also be employed in computing the similarity between two frames.

4.2. Clustering

As described in Table 1, the dominant-set clustering algorithm begins with the similarity matrix and iteratively bipartitions the frames into dominant set and non-dominant set, therefore, produces the clusters progressively and hierarchically. The clustering process usually stops when all frames are grouped into one of the clusters or when certain criteria are satisfied. We choose to terminate the clustering process when more than 90% frames in a shot are clustered so as to avoid forming tiny and meaningless clusters. The rest frames are assigned to the formed clusters or ignored directly as noise using the fast assignment algorithm as shown in Table 2.

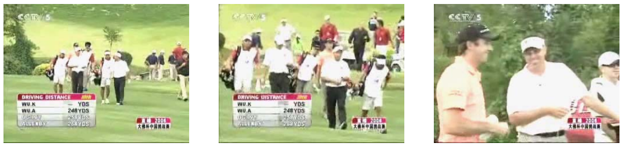
Alternatively, the clustering can be terminated when the maximal number of clusters has been reached. This criteria is particularly useful for embedded devices or systems, such as mobiles, where limited resources and processing time are available.

4.3. Selecting Key Clusters

Some clusters formed by the dominant-set clustering may be not sufficiently significant. A common approach is to discard the clusters whose sizes, i.e. number of frames, are smaller than a threshold. The remaining clusters are considered as key clusters from which key frames are to be extracted. However [3] argues that a key cluster should represent at least one uninterrupted sequence of frames longer than a duration threshold. Since the similarity between two frames does not include the temporal information, the frames in a cluster may not be consecutive in time. In our method, we select the consecutive sequences with a tolerance of several interrupted frames. This strategy has been proved effective in practice. The threshold should be adaptive to the length of the shot. In experiments, it is set to fifteen percent of the shot's length. The clusters containing a sequence of consecutive frames longer than this duration threshold are selected as key clusters while other clusters are simply discarded.

4.4. Selecting Key Frames

In most cases, a key cluster consists of only one consecutive sequence of frames which are long enough and the frame that is closest to the centroid of the key cluster in the feature space is chosen as the key frame. However, for complicated shots, such as those taken during zooming-in and -out operations, it is possible that a key cluster is composed of multiple consecutive sequences of frames which are all long enough. In this case, we select multiple key frames from the cluster, each being the middle frame of the corresponding consecutive long sequence. Notice that we simply choose the middle frame instead of the frame closest to the key cluster center for each sequence in order to avoid the selected key frames being temporally too close.



2211(2152-2430) 2352(2152-2430) 3042(2992-3100)

Fig. 2. Examples of key frames extracted from a golf video

Table 3. Experimental Results

	shots	Key frames	Key-frame /shot	Percentage
News video	210	227	1.08	0.35%
Entertainment	208	326	1.57	0.76%
Home video	269	330	1.23	0.48%
Sports video	138	262	1.90	0.96%

5. EXPERIMENTAL RESULTS

5.1. Experimental Setup and Results

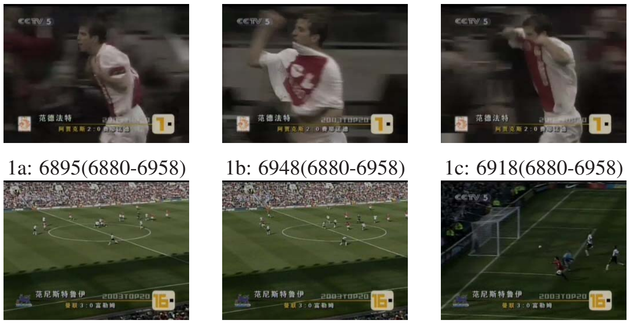
More than three hours of various types of videos including news, entertainment, home and sports are used to evaluate the proposed method. Each video sequence is segmented into shots by the twin-comparison approach [9]. The frames in gradual transition are ignored and shots whose duration are less than a threshold (e.g.,30 frames) are merged with their neighbor shots to avoid meaningless shots. Then key frames are extracted for each shot by employing our proposed method. Figure 2 shows two typical examples: the first two key frames are extracted for a zooming-in golf shot and the third key frame is selected for a static shot focusing on several players. The detailed experiment results are summarized in Table 3, where the *key-frame/shot* ratio represents the averaged number of key frames per shot and the *percentage* stands for the percentage of key frames over the video sequence. It is noticed that sports and entertainment videos have higher *key-frame/shot* ratios and higher *percentages* than news and home videos. This is consistent to the fact that usually more actions and content variations exist in sports and entertainment videos. This result illustrates qualitatively the effectiveness of our method in its dynamic adaption to video content.

5.2. Subjective Evaluation and Comparison

So far, there are no standard and consistent framework to systematically evaluate the performance a key-frame extraction method. This may be partly due to the subjectiveness of the definition of key frames and partly due to the lack of benchmarking databases with ground truth. [1] describes three approaches for evaluating key-frame extraction methods: descriptive evaluation, objective metrics and subjective user study. The descriptive evaluation is inadequate and the objective metrics is often biased toward certain summarization viewpoints. We conduct the subjective user study similar to Liu *et al.* [10]. The key frames extracted from each shot are examined by human subjects and a score out of the three scales representing Good, Acceptable and Bad is assigned to each shot. The score takes the meanings, coverage and re-

Table 4. Evaluation Results

	Middle Frame			Adaptive Clustering			Our Approach		
	Good	Acceptable	Bad	Good	Acceptable	Bad	Good	Acceptable	Bad
News video	82.38%	9.19%	8.37%	88.57%	7.14%	4.29%	90.0%	6.19%	3.81%
Entertainment	41.35%	30.77%	27.88%	62.50%	24.04%	13.46%	81.73%	14.42%	3.85%
Home video	66.78%	18.15%	15.07%	72.26%	20.81%	6.85%	84.93%	11.64%	3.42%
Sports video	43.59%	23.08%	33.33%	56.41%	26.92%	16.67%	74.34%	20.52%	5.13%



1a: 6895(6880-6958) 1b: 6948(6880-6958) 1c: 6918(6880-6958)

2a: 2776(2701-3065) 2b: 2819(2701-3065) 2c: 3002(2701-3065)

Fig. 3. key frames for two shots extracted by adaptive clustering and the proposed method

dundancy of the key frames into consideration. Scores on a large collection of shots by ten subjects are used to verify the performance of a key frame extraction algorithm.

We compare our method with the two commonly used methods. One is to select the middle frame of a shot as its key-frame due to its simplicity and the other is the adaptive clustering proposed in [2] since it is effective, efficient and adaptive to visual content. In the experiments, we set $\delta = 0.85, M = 4$ for the adaptive clustering [2] because in most cases four key frames are sufficient to represent a shot. Note that the parameters, especially δ, are important to the performance of the adaptive clustering. Ten subjects are asked to give scores based on their satisfaction to how well the key frames extracted by three different methods capture the salient content of a shot. Table 4 shows the statistical results for each type of videos. The middle frame approach performs worst for all types of videos because it lacks of adaptation to the content of a shot. Our method and adaptive clustering perform comparably for news video whereas our method performs much better for the rest types of videos. The reason is that the adaptive clustering is very sensitive to the choice of its parameters and, therefore, there hardly exists a set of parameters that work for all types of videos.

In Figure 3, the first row shows key frames for a players close-up shot: (1a) and (1b) extracted by the adaptive clustering are redundant whereas (1c) extracted by our approach is compact. The second row shows key frames for a ball shoot shot: (2a) extracted by the adaptive clustering is less informative than (2b) and (2c) extracted by our approach.

6. CONCLUSIONS

In this paper, we have presented a method for key-frame extraction based on dominant-set clustering. The method is computationally simple and dynamically determines the number of key frames. Experiments on various types of real videos have shown that the method is adaptive to the video content. In addition, the proposed method produces key frames progressively, which is desirable for embedded devices. Furthermore, the method can be employed to both shots and clips, and can be easily extended to any other useful visual features, such as motion, and different distance metrics.

7. ACKNOWLEDGMENT

This work is partly supported by NSFC (Grant No. 60672040, 60705003).

References

[1] B.T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. on Multimedia Computing, Communications and Applications*, vol. 3, no. 1, 2007.

[2] Y. Zhuang, Y. Rui, T.S Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," *in Proc. ICIP*, vol. 1, pp. 866–870, 1998.

[3] A. Girgensohn and J. Boreczky, "Time-Constrained Keyframe Selection Technique," *Multimedia Tools and Applications*, vol. 11, no. 3, pp. 347–358, 2000.

[4] X.D. Yu, L. Wang, Q. Tian, and P. Xue, "Multilevel video representation with application to keyframe extraction," *10th International Multimedia Modelling Conference*, pp. 117–123, 2004.

[5] D. Gibson, N. Campbell, and B. Thomas, "Visual Abstraction of Wildlife Footage using Gaussian Mixture Models and the Minimum Description Length Criterion," *in Proc. ICPR*, vol. 2, pp. 814–817, 2002.

[6] M. Pavan and M. Pelillo, "Efficient Out-of-Sample Extension of Dominant-Set Clusters," *Advances in Neural Information Processing Systems*, vol. 17, pp. 1057–1064, 2005.

[7] M. Pavan and M. Pelillo, "A new graph-theoretic approach to clustering and segmentation," *in Proc. CVPR*, pp. 3895–3900, 2003.

[8] M. Mitra W. Zhu J. Huang, S. R. Kumar and R. Zabih, "Image indexing using Color Correlograms," *in Proc. CVPR*, pp. 762–768, 1997.

[9] H.J. Zhang, A. Kankanhalli, and S.W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, vol. 1, no. 1, pp. 10–28, 1993.

[10] T. Liu, H.J. Zhang, and F. Qi, "A novel video keyframe-extraction algorithm based on perceived motion energy model," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 10, pp. 1006–1013, 2003.