Efficient Video Shot Summarization Using an Enhanced Spectral Clustering Approach

Vasileios Chasanis, Aristidis Likas, and Nikolaos Galatsanos

Department of Computer Science, University of Ioannina,
45110 Ioannina, Greece
{vchasani,arly,galatsanos}@cs.uoi.gr

Abstract. Video summarization is a powerful tool to handle the huge amount of data generated every day. At shot level, the key-frame extraction problem provides sufficient indexing and browsing of large video databases. In this paper we propose an approach that estimates the number of key-frames using elements of the spectral graph theory. Next, the frames of the video sequence are clustered into groups using an improved version of the spectral clustering algorithm. Experimental results show that our algorithm efficiently summarizes the content of a video shot producing unique and representative key-frames outperforming other methods.

Keywords: Video summarization, Key-frame extraction, Spectral clustering, Global k-means.

1 Introduction

Due to the rapid increase in the amount of video data generated nowadays, appropriate video indexing and browsing tools are required to efficiently store videos in large databases. Such indexed databases assist the efficient retrieval of any video in modern video retrieval systems. The most popular indexing and summarization method is based on key-frame extraction. More specifically, a video shot, which is the smallest physical segment of a video (defined as an unbroken sequence of frames recorded from one camera) can be sufficiently summarized using its most representative frames, which are the key-frames. A good summarization of a shot contributes to an efficient indexing and browsing of large video databases. The key-frame extraction problem seeks to find the most representative frames of a shot. However any such algorithm should fulfil some requirements. Firstly, the key-frames should represent the whole video content without missing important information and secondly, these key-frames should not be similar, in terms of video content information, thus containing redundant information.

Several approaches have been proposed for the key-frame extraction problem. We assume that a video is already segmented into shots using any shot detection algorithm. The simplest methods choose the first, last and median frames of a shot or a combination of the previous ones to describe the content of a shot [7].

V. Kůrková et al. (Eds.): ICANN 2008, Part I, LNCS 5163, pp. 847–856, 2008.

In [9] the optical flow is computed and the local minima of a motion metric are selected as key-frames. In [2] it is proposed to form a trajectory from the feature vectors for all frames within a shot. The magnitude of the second derivative of this feature trajectory with respect to time is used as a curvature measure in this case. As key-frames the local minima and maxima of this magnitude are selected. In [12] multiple frames are detected using unsupervised clustering based on the visual variations in shots. A main drawback of this algorithm is the determination of the appropriate number of key-frames to represent each shot which depends on the threshold parameter that controls the density of the clusters. A variant of this algorithm is presented in [6] where the final number of key-frames depends on a threshold parameter which defines whether two frames are similar. In [3] the extraction of the key-frames is based on detecting curvature points within the curve of cumulative frame differences and suggest quality measures to evaluate the summaries.

In our approach the frames of a video sequence are clustered into groups using an improved version of the typical spectral clustering method [5] that uses the global k-means algorithm [4] in the clustering stage after the eigenvector computation. Moreover, the number of key-frames is estimated using results from spectral graph theory. The rest of the paper is organized as follows: In section 2 we present our key-frame extraction algorithm. In section 3 a method for the estimation of the number of clusters is proposed and in section 4 we present quality measures for the evaluation of the summary of a shot. In section 5 we present numerical experiments and compare our method with three other approaches. Finally in section 6 we provide conclusions and suggestions for further study.

2 Key-Frames Extraction

To deal with the problem of key frame extraction two major issues must be addressed. Firstly the number of key-frames must me estimated and second an efficient algorithm must produce unique key-frames that will summarize the content of the shot. In our work for each frame a 16-bin HSV normalized histogram is used, with 8 bins for hue and 4 bins for each of saturation and value.

2.1 The Typical Spectral Clustering Algorithm

To perform key-frame extraction the video frames of a shot are clustered into groups using an improved spectral clustering algorithm. Then, the medoid of each group, defined as the frame of a group whose average similarity to all other frames of this group is maximal, is characterized as a key-frame. The main steps of the typical spectral clustering algorithm [5] are described next. Suppose there is a set of objects $S = s_1, s_2, \ldots, s_N$ to be partitioned into K groups.

1. Compute similarity matrix $A \in \mathbb{R}^{N \times N}$ for the pairs of objects of the data set S.
2. Define D to be the diagonal matrix whose (i, i) element is the sum of the A's i-th row and construct the Laplacian matrix $L = I - D^{-1/2} A D^{-1/2}$.

3. Compute the K principal eigenvectors $x_1, x_2, \ldots, x_K$ of matrix L to build an $N \times K$ matrix $X = [x_1\ x_2 \ldots\ x_K]$.
4. Renormalize each row of X to have unit length and form matrix Y so that:

$$y_{ij} = x_{ij}/(\sum_j x_{ij}^2)^{1/2}\ . \tag{1}$$

5. Cluster the rows of Y into K groups using k-means.
6. Finally, assign object s_i to cluster j if and only if row i of the matrix Y has been assigned to cluster j.

In what concerns our key-frame extraction problem, suppose we are given a data set $H = H_1, \ldots, H_N$ where H_n is the feature vector (normalized color histogram) of the n-th frame. The distance function we consider is the Euclidean distance between the histograms of the frames. As a result each element of the similarity matrix A is computed as follows:

$$a(i,j) = 1 - \sqrt{\sum_{h \in bins} (H_i(h) - H_j(h))^2}\ . \tag{2}$$

2.2 Global k-Means

Furthermore, in the fifth step of the spectral clustering algorithm instead of using the typical k-means approach, we have used the very efficient global k-means algorithm [4]. Global k-means in an incremental deterministic clustering algorithm that overcomes the important initialization problem of the typical k-means approach. This initialization problem has been found to be severe in the case of frame clustering, significantly affecting the quality of the key-frames. Using the global k-means, the obtained key frames usually provide a sensible representation of shot content. Next we briefly review the global k-means algorithm. Suppose we are given a data set $X = x_1, \ldots, x_N$, $x_n \in R^d$ to be partitioned into K disjoint clusters $C_1, C_2, \ldots, C_K$.

This algorithm is incremental in nature. It is based on the idea that the optimal partition into K groups can be obtained through local search (using k-means) starting from an initial state with i) the k-1 centers placed at the optimal positions for the (k-1)-clustering problem and ii) the remaining k-th center placed at an appropriate position within the dataset. Based on this idea, the K-clustering problem is incrementally solved as follows. Starting with k=1, find the optimal solution which is the centroid of the data set X. To solve the problem with two clusters, the k-means algorithm is executed N times (where N is the size of the data set) from the following initial positions of the cluster centers: the first cluster center is always placed at the optimal position for the problem with k=1, whereas the second center at execution n is initially placed at the position of data x_n. The best solution obtained after the N executions of k-means is considered as the solution for k=2. In general if we want to solve the problem with k clusters, N runs of the k-means algorithm are performed, where each run n starts with the k-1 centers initially placed at the positions corresponding to the solution obtained for the

(k-1)-clustering problem, while the k-th center is initially placed at the position of data x_n. A great benefit of this algorithm is that it provides the solutions for all k-clustering problems with $k \leq K$.

3 Estimation of Number of Key-Frames Using Spectral Graph Theory

As already mentioned in the introduction, the number of key-frames cannot be predetermined due to the different content of each shot. In our approach we attempt to estimate the number of the key-frames using results from the spectral graph theory.

Assume we wish to partition dataset S into disjoint subsets $(S_1, \ldots, S_K)$, and let $X = [X_1, \ldots, X_K] \in \mathbb{R}^{N \times K}$ denote the partition matrix, where X_j is the binary indicator vector for set S_j such that:

$$\begin{array}{l} X(i,j) = 1 \ : \ if \ i \in S_j \\ X(i,j) = 0 \ : \ otherwise \end{array} . \tag{3}$$

The optimal solution [10] is defined as:

$$\begin{array}{c} \max\limits_{X} \ trace(X^T L X) \\ s.t. \ X^T X = I_K \text{ and } X(i,j) \in \{0,1\} \end{array} , \tag{4}$$

where L is the Laplacian matrix defined in section 2.1. The spectral clustering algorithm (for K clusters) provides solution to the following optimization problem:

$$\begin{array}{c} \max\limits_{Y} \ trace(Y^T L Y) \\ s.t. \ Y^T Y = I_K \end{array} . \tag{5}$$

Relaxing Y into the continuous domain turns the discrete problem into a continuous optimization problem. The optimal solution is attained at $Y = U_K$, where the columns u_i of $U_k, i = 1, \ldots, K$, are the eigenvectors corresponding to the ordered top K largest eigenvalues λ_i of L. Since it holds that [11]:

$$\lambda_1 + \lambda_2 + \ldots + \lambda_K = \max_{Y^T Y = I_K} trace(Y^T L Y) \ , \tag{6}$$

the optimization criterion that also quantifies the quality of the solution for K clusters and its corresponding difference for successive values of K are respectively given by:

$$\begin{array}{l} sol(K) = \lambda_1 + \lambda_2 + \ldots + \lambda_K \\ sol(K+1) - sol(K) = \lambda_{K+1} \end{array} . \tag{7}$$

When the improvement in this optimization criterion (i.e. the value of the λ_{K+1} eigenvalue) is below a threshold, improvement by the addition of cluster K+1

is considered negligible, thus the estimate of the number of clusters is assumed to be K. The threshold value that is used in all our experiments was fixed to Th=0.005 with very good results.

4 Summary Evaluation

A difficult issue of the key-frame extraction problem is related to the evaluation of the extracted key-frames, since it is rather subjective which frames are the best representatives of the content of a shot. There are several quality measures that can be used to evaluate the efficiency of the algorithms. In [3], two quality measures are used. The first is the Fidelity measure proposed in [1] and the second is the Shot Reconstruction Degree measure proposed in [8].

4.1 Average Fidelity

The Fidelity measure compares each key-frame with other frames in the shot. Given the frame sequence $F = \{F_1, F_2, \ldots, F_N\}$ and the set of key-frames $KF = \{KF_1, KF_2, \ldots, KF_{N_{kf}}\}$ the distance between the set of key-frames KF and a frame F_n is defined as:

$$d(F_n, KF) = \min_j \; Diff(F_n, KF_j),\; j = 1, 2, \ldots, N_{kf} \;, \tag{8}$$

where N_{kf} is the number of key-frames and $Diff(F_i, F_j)$ a distance measure between two frames F_i and F_j. The Fidelity measure is computed as:

$$\text{Fidelity}(F, KF) = \text{MaxDiff} - d_{all}(F, KF) \;, \tag{9}$$

where MaxDiff is a constant representing the largest possible value that the frame difference measure can assume and $d_{all}(F, KF)$ is as follows:

$$d_{all}(F, KF) = \max_n \; d(F_n, KF),\; n = 1, 2, \ldots, N \;. \tag{10}$$

However as mentioned in [8], Fidelity cannot capture well the dynamics of a shot since it focuses on global details. For that reason we compute the Average Fidelity which is computed using the average of the minimal distances between the key frame set and the video shot and is given from the following equation:

$$\text{Average Fidelity}(F, KF) = \text{MaxDiff} - \frac{1}{N} \sum_{n=1}^{N} d(F_n, KF) \;. \tag{11}$$

4.2 Shot Reconstruction Degree

Given the set of key-frames, the whole frame sequence of a shot can be reconstructed using an interpolation algorithm. The better the reconstructed video sequence approximates the original sequence, the better the set of key-frames

summarizes the video content. More specifically, given the frame sequence F, the set of key-frames KF and a frame interpolation algorithm IA(), we can reconstruct any frame from a pair of key-frames in KF [8]:

$$\tilde{F}_n = \text{IA}(KFn_j, KFn_{j+1}, n, n_j, n_{j+1}),\ n_j \leq n < n_{j+1}\ . \tag{12}$$

The Shot Reconstruction Degree (SRD) measure is defined as follows:

$$SRD(F, KF) = \sum_{n=0}^{N-1} (Sim(F_n, \tilde{F}_n))\ , \tag{13}$$

where Sim() is given from the following equation:

$$Sim(F_n, \tilde{F}_n) = \log(MaxDiff/Diff(F_n, \tilde{F}_n))\ , \tag{14}$$

where $Diff(F_i, F_j)$ is a distance measure between two frames F_i and F_j and MaxDiff the largest possible value that the frame difference measure can assume.

5 Experiments

In this section we present the application and evaluation of our key-frame extraction algorithm compared to three other algorithms.

5.1 Data

In our experiments we used seven frame sequences. The first frame sequence describes an action of a comedy movie that takes place in an office. The other six sequences are taken from sports. Three of them describe three attempts in the NBA Slam Dunk Contest and the other three a goal attempt in a football match taken from three individual cameras. In *Table 1* we present the characteristics of the video data set.

5.2 Comparison

We compare the proposed approach to three other methods. The first one is the simple k-means algorithm. For each shot we perform 20 iterations of the k-means algorithm keeping the iteration with the minimum clustering error. The number of clusters in k-means algorithm is assumed to be the same with one selected

Table 1. Characteristics of video data set

	FRAME SEQUENCE						
	F_1	F_2	F_3	F_4	F_5	F_6	F_7
No. FRAMES	633	144	145	146	225	300	172
GENRE	Comedy	Basketball	Basketball	Basketball	Football	Football	Football

using the proposed estimation algorithm of section 3. The second technique is presented in [6], as a variant of the method presented in [12]. Initially, the middle frame of the video sequence is selected as the first key-frame and added to the empty set of key-frames KF. Next, each frame in the video sequence is compared with the current set of key-frames. If it differs from every key-frame in the current set, then it is added into the set as a new key-frame. This algorithm uses a threshold to discriminate whether two frames are similar or not. In our experiments this threshold parameter is set to such a value that the number of key-frames extracted is the same as in our algorithm. Finally, the third technique is the spectral clustering algorithm employing the simple k-means algorithm.

5.3 Evaluation

To evaluate the results of the extracted key-frames we use the metrics mentioned in section 4. More specifically in *Tables 2-3* we present the performance results for the Average Fidelity and SDR measures respectively. To compute the SDR we use a simple linear interpolation algorithm on the frame's features. It is clear that our approach provides the best summarization of each shot compared to the other methods and the best reconstruction of the original video sequence from the extracted key-frames.

Table 2. Comparative results of the tested key-frame extraction algorithms using Average Fidelity

AV.FIDELITY	FRAME SEQUENCE						
ALGORITHM	F_1	F_2	F_3	F_4	F_5	F_6	F_7
Our method	0.973	0.9437	0.9507	0.9559	0.9687	0.9546	0.978
K-means	0.9549	0.9278	0.9344	0.948	0.9467	0.931	0.9654
Method in [6]	0.9616	0.8913	0.9268	0.9405	0.955	0.9424	0.9672
Typical Spectral Algorithm	0.9619	0.9235	0.9253	0.9462	0.9625	0.9318	0.9675

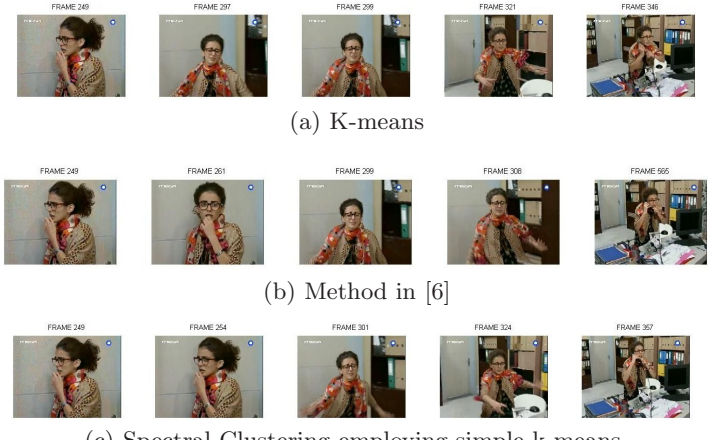
Table 3. Comparative results of the tested key-frame extraction algorithms using SDR

SDR	FRAME SEQUENCE						
ALGORITHM	F_1	F_2	F_3	F_4	F_5	F_6	F_7
Our method	1859.66	425.87	511.58	527.32	855.96	860.01	711.18
K-means	1533.34	369.87	430.78	356.46	808.24	753.75	648.71
Method in [6]	1693.1	292.43	374.23	340.89	758.23	813.1	642.97
Typical Spectral Algorithm	1620.6	362.64	431.32	393.02	780.33	791.2	663.15

5.4 Representation

As already mentioned in section 3.2 a great benefit of the global k-means algorithm is that it provides the solutions for all intermediate k-clustering problems with $k \leq K$. In *Fig. 1* we give an example of the extracted key-frames of a video shot with object and camera motion. Moving from the top to the bottom of

Fig. 1. Key-frame extraction of a shot using the proposed method($N_{kf} = 5$)



(a) K-means

(b) Method in [6]

(c) Spectral Clustering employing simple k-means

Fig. 2. Results for the key-frame extraction algorithms used for comparison

this figure we show all solutions until the selected number of key-frames N_{kf}=5 is reached. The shot that we used contains 633 frames. It shows a woman in an office set-up. This shot can be semantically divided into 5 sub-shots. a) The woman stands against a door eavesdropping and then rushes to her office to pick up the phone that is ringing; b) she talks on the phone, c) lays the receiver of the phone down with a visible effort not to make any noise, d) she rushes back to the door, and e)she continues eavesdropping. In *Fig. 2* we provide the key-frames extracted performing the simple k-means algorithm, the algorithm in [6] and the typical spectral clustering algorithm. All algorithms fail to provide a solution

(a) Our method

(b) K-means

(c) Method in [6]

(d) Spectral Clustering employing simple k-means

Fig. 3. Key-frame extraction algorithms in comparison in basketball sequence

adequately describing the visual content of the shot, whereas our approach provides a sensible solution. More specifically, they do not produce any frames for sub-shots (c), (d) and (e) and instead produce multiple frames for sub-shot (a). In contrast the proposed approach produces key frames for all sub-shots. In *Fig. 3* we provide the key-frames for these four algorithms for a video shot containing 145 frames and describing a slam dunk attempt. It becomes clear that our algorithm summarizes the attempt from the beginning to the end, whereas the other three fail to describe the end of the action.

6 Conclusions

In this paper a new method for key-frame extraction is proposed. Key-frames are extracted using a spectral clustering method employing the global k-means algorithm in the clustering procedure. Furthermore, the number of key-frames is estimated using results from the spectral graph theory, by examining the eigenvalues of the similarity matrix corresponding to pairs of shot frames. Appropriate quality measures indicate that our method outperforms traditional techniques and provides efficient summarization and reconstruction of a video sequence from the extracted key-frames. In future work, we will try to improve the performance of our method by examining other features for the frames, such as motion and edge histograms.

Acknowledgments

This research project (PENED) is co-financed by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%).

References

1. Chang, H.S., Sull, S., Lee, S.U.: Efficient Video Indexing Scheme for Content-Based Retrieval. IEEE on Transactions Circuits and Systems Video Technology 9(8), 1269–1279 (1999)
2. Doulamis, A.D., Doulamis, N.D., Kollias, S.D.: Non-sequential video content representation using temporal variation of feature vectors. IEEE Transactions on Consumer Electronics 46(3), 758–768 (2000)
3. Gianluigi, C., Raimondo, S.: An innovative algorithm for key frame extraction in video summarization. Journal of Real-Time Image Proceesing 1(1), 69–88 (2006)
4. Likas, A., Vlassis, N., Verbeek, J.J.: The global k-means clustering algorithm. Pattern Recognition 36(2), 451–461 (2003)
5. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Proceedings Neural Information Processing Systems (NIPS 2001) (2001)
6. Rasheed, Z., Shah, M.: Detection and representation of scenes in videos. IEEE Transactions on Multimedia 7(6), 1097–1105 (2005)
7. Rui, Y., Huang, T.S., Mehrotra, S.: Exploring video structure beyond the shots. In: Proceedings of IEEE International Conference on Multimedia Computing and Systems (ICMCS), Texas, USA, pp. 237–240 (1998)
8. Tieyan, L., Zhang, X., Feng, J., Lo, K.T.: Shot reconstruction degree: a novel criterion for key frame selection. Pattern Recognition Letters 25, 1451–1457 (2004)
9. Wolf, W.: Key frame selection by motion analysis. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1228–1231 (1996)
10. Xing, E.P., Jordan, M.I.: On semidefinite relaxation for normalized k-cut and connections to spectral clustering. Technical Report CSD-03-1265, Computer Science Division, University of California, Berkeley (2003)
11. Zha, H., Ding, C., Gu, M., He, X., Simon, H.: Spectral relaxation for k-means clustering. In: Neural Information Processing Systems (NIPS 2001) (2001)
12. Zhuang, Y., Rui, Y., Huang, T.S., Mehrotra, S.: Adaptive key frame extraction using unsupervised clustering. In: Proceedings of IEEE International Conference on Image Processing, pp. 866–870 (1998)