An Agglomerative Approach for Shot Summarization Based on Content Homogeneity

Antonis Ioannidis, Vasileios Chasanis and Aristidis Likas
Department of Computer Science and Engineering, University of Ioannina, Greece

ABSTRACT

An efficient shot summarization method is presented based on agglomerative clustering of the shot frames. Unlike other agglomerative methods, our approach relies on a cluster merging criterion that computes the content homogeneity of a merged cluster. An important feature of the proposed approach is the automatic estimation of the number of a shot's most representative frames, called keyframes. The method starts by splitting each video sequence into small, equal sized clusters (segments). Then, agglomerative clustering is performed, where from the current set of clusters, a pair of clusters is selected and merged to form a larger unimodal (homogeneous) cluster. The algorithm proceeds until no further cluster merging is possible. At the end, the medoid of each of the final clusters is selected as keyframe and the set of keyframes constitutes the summary of the shot. Numerical experiments demonstrate that our method reasonable estimates the number of ground-truth keyframes, while extracting non-repetitive keyframes that efficiently summarize the content of each shot.

Keywords: Summarization, keyframe extraction, unimodality test, agglomerative clustering.

1. INTRODUCTION

In recent years, high quality mobile devices like notepads, smartphones and tablets, along with already existed digital cameras produce enormous amounts of digital video. As a result, in order to be able to manage such a volume of video data, new applications are developed aiming for better storage, indexing and video retrieval. A video usually contains a large amount of frames, which is difficult to process for several applications. Keyframe extraction techniques try to provide sufficient summarization of video shots, which allows the user to have a fast, descriptive and compact knowledge of a shot content. There are two major categories of key-frame extraction methods. The first is based on the detection of abrupt changes in the similarity between successive frames [1]. In [2], three properties (Iso-Content Distance, Iso-Content Error and Iso-Content Distortion) are considered. The selected keyframes are equidistant in the shot content curve with respect to those properties. This type of keyframe extraction has the disadvantage that it may extract similar keyframes, if the same content reappears during a shot.

Our method belongs to the second major category of key-frame extraction methods that consider keyframe extraction as a clustering problem, thus frames are clustered into groups and the cluster representatives (e.g. medoids) are selected as keyframes. For example, in [3] the keyframes are extracted using clustering based on the visual variation in a shot. A variant of this algorithm is presented in [4], where a threshold parameter defining whether two frames are similar, controls the final number of key-frames. In [5], frames are clustered into groups with a split-merge approach based on mutual information. A different technique for key-frame selection is described in [6], where the position of key-frames in the sequence is also taken into account. In [7] the number of key-frames is estimated using elements of the spectral graph theory. Next, the frames sequence is clustered into groups using an improved spectral clustering algorithm.

In this paper, we propose an agglomerative clustering approach for shot summarization via keyframe extraction that is based on cluster unimodality and automatically estimates the number of shot keyframes. Initially, a shot sequence is splitted into small, equal sized clusters (segments). Then, the method iteratively merges clusters aiming to form larger unimodal clusters. In order to decide whether a cluster is unimodal or not, the dip-dist criterion [8] is used. The test for unimodality of a cluster essentially decides on the content homogeneity of the frames in this cluster. In this way, shot frames that belong to the same cluster, present similar visual content. Cluster merging proceeds until there is no merging providing a unimodal outcome. After clustering, the medoid of each formed group of frames is selected as keyframe. Note that our algorithm provides unique key-frames in the sense that, if the same frame subsequence appears several times during a shot, only one key-frame will be selected as representative for the subsequence. Moreover, in the final clustering solution, a cluster does not necessarily contain frames that are adjacent in the shot frame sequence. The rest of the paper is organized as follows. In Section 2 we describe the statistical test for unimodality used in our method. In Sections 3 and 4 we present the criterion for accepting or rejecting a merge of two clusters and the details of the proposed agglomerative approach for key-frame extraction. In Section 5 we describe the evaluation procedure and provide comparative experimental results. Finally, in Section 6 we provide conclusions and suggestions for further study.

2. DECIDING ON CLUSTER UNIMODALITY

The proposed method relies on a unimodality test to decide on the homogeneity of a set of frames. To check for unimodality of a set of frames we employ *the dip-dist criterion* [8], proposed for a evaluating the cluster structure of a set of data objects. This criterion is based on Hartigan's dip-test [9] for unimodality. The basic intuition behind dip-dist is that if the density distribution of a set of objects is unimodal, then the set is considered homogeneous. However, unimodality is not checked in the original data space, but it is tested using only the pairwise distances between data objects (i.e. the distance matrix).

More specifically, to compute the dip-dist criterion [8], for a set of objects (frames), each object (frame) of the set is treated as *viewer* that decides on the unimodality of the set by considering the set of the pairwise distances from the viewer to all other data objects (frames). Then, the density of this set of distances is tested for unimodality using Hartigans' dip test [9] and is characterized as either unimodal or multimodal. If the percentage of viewers suggesting multimodality exceeds a given threshold, then the set of objects is characterized as multimodal, otherwise it is considered unimodal. As proposed in [9], given a set of *n* real values F_n, the dip-test computes the *dip value* of F_n ($dip(F_n)$) which is the *departure from unimodality* of the empirical distribution (cdf) of F_n:

$$dip(F_n) = \min_{G\in\mathbb{U}} \rho(F_n, G), \tag{1}$$

where $\rho(F_n,G)$ is an appropriately defined distance between the two distributions F_n, G and $\mathbb{U}$ the class of all unimodal distributions. An efficient algorithm is proposed in [9] to compute $dip(F_n)$. It is also argued that uniform distribution U is the most appropriate for the null hypothesis. Thus, if $dip(F_n)$ is the dip statistic for a set F_n, the computation of the *p*-value for this unimodality test uses *b* bootstrap sets U_n^r, r=1,…, b (each containing *n* samples from U [0,1]) and expresses the probability of $dip(F_n)$ being less than $dip(U_n^r)$.

$$P = \#[dip(F_n) \le dip(U_n^r)]/b, r = 1,\ldots,b. \tag{2}$$

The null hypothesis H_0 that F_n is unimodal, is accepted at significance level α if $P>\alpha$, otherwise hypothesis H_1 indicating multimodality, is accepted.

3. MERGING A PAIR OF CLUSTERS

The previously described dip-dist criterion for deciding on the unimodality of a set (cluster) of frames can be exploited to develop an agglomerative for clustering the frames of a shot: we divide the shot into small segments to get initial small clusters and proceed with iteratively merging smaller clusters into larger ones taking into account the unimodality constraint. This means that in order to accept the merge of two clusters, the resulting union cluster should remain unimodal. The test for cluster merging is specified below:

Suppose we are given a video shot sequence with *N* frames and their corresponding features, HSV color histograms in our case. The sequence is initially divided into equal-sized small segments, called *s-clusters*. The size *L* of *s-clusters* is specified by the user. In this way, the shot sequence is segmented into $M=N/L$ successive, non-overlapping *s-clusters*, which form the initial clustering $S_0= \{s_1,\ldots, si,.., s_M\}$, i=1,…,M.

In order to form larger unimodal clusters, we iteratively merge *s-clusters*. In order to select the pair of *s-clusters* that will be merged, the unimodality test is applied to every pair of *s-clusters*. More specifically, to test for the unimodality of a pair of two *s-clusters* s_i and s_j, we temporarily merge them into one cluster $s_i \cup s_j$ and apply the dip-dist criterion to decide whether this cluster is unimodal or not, i.e. if its visual content, as specified by the frames descriptors, is homogeneous or not. If the dip-dist criterion decides multimodality, we consider that the visual content changes in this cluster, thus *s-clusters* s_i and s_j should not be merged, since their content is different. More specifically, based on the description in the previous section, if at least one frame/viewer of cluster $s_i \cup s_j$ suggests multimodality, (it is called *split-viewer* [8]), then *s-clusters* s_i and s_j cannot not be merged, otherwise they could be merged. To show the details of dip-dist computation, suppose that a temporarily formed union cluster $s_i \cup s_j$ contains the *2L* frames $f=\{f_1, ..., f_{2L}\}$ with $Vf=\{Vf_1, ..., Vf_{2L}\}$, where Vf_i be the feature vector (image descriptor) of the corresponding frame. We first form the *2L*x*2L* matrix *Dist* with the pairwise Euclidean distances of the frames:

$$Dist_{ij} = \| Vf_i - Vf_j \|_2 . \tag{3}$$

Then the dip-dist criterion used to decide whether cluster $s_i \cup s_j$ is unimodal or not, is applied as follows:

1. Create *b* sets U_{2L}^r of *2L* values sampled form *U*(0,1) and compute the dip values $dip(U_{2L}^r)$, *r*=1,…,*b*, for those sets. It is worth mentioning that this can be done in a preprocessing step and the obtained dip values can be stored and used in all other dip-dist computations.

2. Compute the dip values *dip(i)* (Eq. 1) for every frame/ viewer f_i, *i*=1,…,*2L* using the values of the *i*-th row of matrix *Dist*.

3. Estimate the *p*-values *P(i)*, *i*=1,…,*2L*, based on Eq. 2 using a significance level α and the percentage of frames/viewers identifying multimodality. In our case we set α=0. If at least one viewer f_i observes multimodality (*p*-value>0), then cluster $s_i \cup s_j$ is characterized as multimodal. Otherwise, cluster $s_i \cup s_j$ is characterized as unimodal. In the latter case, we compute the *average dip statistic* (*ADS*) of the cluster by considering the dip values of all its members as follows:

$$ADS(s_i \cup s_j) = \frac{1}{2L}\sum_{k=1}^{2L} Dip(f_k). \tag{4}$$

   ADS can be considered as an indication of the "degree" of unimodality. The lower the *ADS* value, the more unimodal the cluster. The *ADS* values are used in every step of our method to determine the pair of clusters that will be actually merged.

The *ADS* values are stored in an *M*x*M* matrix *DS* that provides information about the unimodality of every pair of *s-clusters* of the initial clustering S_0 as follows:

$$DS(i,j) = \begin{cases} ADS(s_i \cup s_j) & if\,(s_i \cup s_j) \text{ is unimodal} \\ 0 & if\,(s_i \cup s_j) \text{ is multimodal} \end{cases}. \tag{5}$$

4. AGGLOMERATIVE CLUSTERING FOR KEYFRAME EXTRACTION

Suppose that the initial set $S=S_0$ of *s-clusters* has been defined and the corresponding values of the *DS* matrix have been computed. As mentioned previously the *ADS* values provide the degree of unimodality of a merged pair of clusters (the lower the better). Thus, in the next step of our approach, it is reasonable to select the *unimodal* pair of clusters s_i, $s_j \in S$ (with *DS(i,j)*>0) having the lowest *ADS* value and merge the corresponding *s-clusters*.

$$(\hat{s}_i, \hat{s}_j) = \underset{s_i, s_j \in S}{\text{argmin}}\, DS(i,j) \quad \text{subject to}\, DS(i,j) > 0. \tag{6}$$

In this way a new cluster $\hat{s}_i \cup \hat{s}_j$ is formed, called *m-cluster*, that replaces the two original *s-clusters* $\hat{s}_i$ and $\hat{s}_j$ in the solution set *S*. To proceed in the following step of our method we have to define the *ADS* value between two *m-clusters*. In case the pair ($m_i \cup m_j$) is found multimodal, using the method described in Section 3, then we set *ADS*($m_i \cup m_j$)=0. Otherwise, if the pair ($m_i \cup m_j$) is found unimodal in order to compute the degree of unimodality *ADS* ($m_i \cup m_j$) another prerequisite must be met. Since the *m-clusters* consist of elementary *s-clusters*, at least one of all the corresponding *s-cluster* pairs ($s_k \cup s_l$) (with $s_k \in m_i$ and $s_l \in m_j$) must be unimodal. Then, the degree of unimodality *ADS* ($m_i \cup m_j$) is given from the following equation:

$$ADS(m_i \cup m_j) = \underset{s_k \in m_i, s_l \in m_j}{\text{argmin}}\, DS(k,l) \quad \text{subject to}\, DS(k,l) > 0. \tag{7}$$

If all the corresponding *s-cluster* pairs ($s_k \cup s_l$) (with $s_k \in m_i$ and $s_l \in m_j$) are multimodal (i.e. *DS(k,l)*=0), then *ADS*($m_i \cup m_j$)=0. All the corresponding *s-cluster* pairs of two *m-clusters* are tested for unimodality, because the union of a very large cluster with a smaller is usually found unimodal since the smaller cluster cannot significantly change the structure of the larger cluster. Thus, the smaller cluster should be unimodal with at least one segment of the larger cluster. Therefore, at each step of our method if the current solution contains *r* clusters *S*= $\{c_1, \ldots, c_r\}$, the pair of clusters $\hat{c}_i$, $\hat{c}_j \in S$ to be merged is decided as:

$$(\hat{c}_i, \hat{c}_j) = \underset{c_i, c_j \in S,}{\operatorname{argmin}} ADS(c_i \cup c_j) \quad \text{subject to}\; ADS(c_i \cup c_j) > 0. \tag{8}$$

The method proceeds until no further merging can be done, thus obtaining the *final clusters*. Finally, the medoid frame of each *final cluster* is selected as key-frame. Actually, the only parameter of our algorithm is the initial size *L* of the *s-clusters*.

5. EXPERIMENTS

In our experiments we have processed 40 video shots with different visual content including car motion, construction demolition, car accidents, changing traffic lights, indoor movement, movie production etc. In total, 30076 frames were processed and 168 ground-truth keyframes were extracted. Two persons have visually extracted keyframes that according to their opinion represent adequately and sufficiently the content of the examined video shots. Most of the video shot sequences have been taken from TRECVID [10].

5.1 Evaluation and Performance Measures

The evaluation of keyframe extraction algorithms is a tedious task because it highly depends on the subjectivity of the person(s) who decide the ground truth keyframes. To evaluate the performance of the presented shot summarization algorithm we use two different measures. The first evaluation measure correlates the clustering solution with the ground-truth keyframes. More specifically, we first assign ground truth keyframes into extracted clusters using the nearest neighbor criterion. Let *G* be the number of the extracted clusters. We form a histogram vector *K* with each member K_j, *j*=1,.., *G* indicating the number of ground-truth keyframes that have been assigned to cluster *j*. Obviously, the histogram vector K^p of a perfect solution would have $K^p{}_j$=1 for all *j*, suggesting an 1-to-1 connection between ground-truth keyframes and extracted clusters. It is obvious that the closer the distribution *K* is to K^p, the better the results. Thus, we define M_1 performance measure, where values closer to zero indicate better performance.

$$M_1 = \frac{1}{G} \| K - K^p \|_2 \ . \tag{9}$$

The second performance evaluation measure, similar to [11] relies on visual comparison of the extracted keyframes and the ground truth keyframes. Two persons with video processing background took part in the evaluation process and the cross-section of their evaluations was used. To evaluate the performance of the proposed algorithm and the algorithms under comparison, we have used F_1 measure [7]. As false keyframe we define the case where a ground-truth keyframe is found similar to more than one of the extracted keyframes.

5.2 Experimental Results

In our experiments, HSV normalized color histograms have been employed to represent shot frames, with 8 bins for hue and 4 bins for each of saturation and value resulting in a 128 (8x4x4) dimensional feature vector. We set the size of the initial clusters *L* equal to 15 frames, the number of bootstraps *b* is set to 1000, while we set *α*=0 for the significance level of dip test. In Table 1 we provide comparative results of our method with clustering methods proposed in [7] and [8]. The method described in [7] is based on spectral clustering and eigenvalue analysis of the distance matrix of the frames. The method in [8] proposes the dip-dist criterion for clustering but applies the criterion in an incremental way (through cluster splitting), while our method is based on cluster merging. For both methods mentioned above, several parameter values have been considered and the best results and presented. Note that for all three compared methods, HSV color histograms have been used and the medoids of the final clusters have been selected as keyframes. As it can be observed, our method is superior in both measures providing non-repetitive keyframes that capture more efficiently the visual content of the shots.

Table 1. Comparative Results using M_1 and F_1 measures.

Method	M_1	F_1
Method in [8]	0.45	73.5
Method in [7]	0.32	75.7
Proposed method	0.30	80.2

6. CONCLUSION

In this paper we have presented an efficient shot summarization method that automatically estimates the number of keyframes to be extracted. The method is based on agglomerative clustering and at each step clusters are iteratively merged to form larger unimodal clusters. In this way clusters are merged only if their union results in a homogeneous larger group of frames. In this way, shot frames that belong to the same cluster are expected to present similar visual content. In order to decide whether a cluster is unimodal or not, the dip-dist criterion has been used that is based on the notion of cluster viewer and employs the Hardigans' dip test for unimodality. The algorithm proceeds until no further merging of clusters is possible. The medoid of each of the cluster is selected as keyframe. Performance results on several video sequences indicate that our method can efficiently estimate the correct number of extracted keyframes while providing non-repetitive keyframes that summarize the shot content. In future work, we plan to test our method using other feature descriptors (motion, SIFT etc.) as well as combinations of them.

ACKNOWLEDGMENT

The work described in this paper is co-financed by the European Regional Development Fund (ERDF) (2007-2013) of the European Union and National Funds (Operational Programme «Competitiveness and Entrepreneurship» (OPCE II), ROP ATTICA), under the Action ``SYNERGASIA (COOPERATION) 2009".

REFERENCES

[1] Gianluigi Ciocca and Raimondo Schettini, “An innovative algorithm for key frame extraction in video summarization.,” Journal of Real-Time Image Processing, vol. 1, no. 1,69–88, (2006).

[2] Costas Panagiotakis, Anastasios D. Doulamis, and Georgios Tziritas, “Equivalent key frames selection based on iso-content principles.,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 19, no. 3, 447–451, (2009).

[3] Yueting Zhuang, Yong Rui, T.S. Huang, and S. Mehrotra, “Adaptive key frame extraction using unsupervised clustering,” in IEEE International Conference on Image Processing, vol. 1, 866 –870, (1998).

[4] Zeeshan Rasheed and Mubarak Shah, “Detection and representation of scenes in videos.,” IEEE Transactions on Multimedia, vol. 7, no. 6, 1097–1105, (2005).

[5] Z. Cernekova, I. Pitas, and C. Nikou, “Information theory-based shot cut/fade detection and video summarization,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 1, 82–91, (2006).

[6] Andreas Girgensohn and John S. Boreczky, “Timeconstrained keyframe selection technique.,” Multimedia Tools Applications, vol. 11, no. 3, 347–358, (2000).

[7] V. Chasanis, A. Likas, and N. Galatsanos, “Scene detection in videos using shot clustering and sequence alignment,” IEEE Transactions on Multimedia, vol. 11, no. 1, 89–100, (2009).

[8] Argyris Kalogeratos and Aristidis Likas, “Dip-means: an incremental clustering method for estimating the number of clusters.,” in Proceedings of the Neural Information Processing Systems, 2012, 2402– 2410, (2012).

[9] J. A. Hartigan and P. M. Hartigan, “The dip test of unimodality,” The Annals of Statistics, vol. 13, no. 1, (1985).

[10] Alan F. Smeaton, Paul Over, and Wessel Kraaij, “Evaluation campaigns and trecvid,” in Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, 321–330, (2006).

[11] Genliang Guan, Zhiyong Wang, Shiyang Lu, Jeremiah Da Deng, and David Dagan Feng, “Keypointbased keyframe selection,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no 4, 729–734, (2013)