Integration of Global and Local Information in Videos for Key Frame Extraction

Dianting Liu[1], Mei-Ling Shyu[1], Chao Chen[1], Shu-Ching Chen[2]
[1]*Department of Electrical and Computer Engineering*
University of Miami, Coral Gables, FL 33124, USA
[2]*Distributed Multimedia Information Systems Laboratory*
School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA
d.liu4@umiami.edu, shyu@miami.edu, c.chen15@umiami.edu, chens@cs.fiu.edu

Abstract

Key frame extraction methods aim to obtain a set of frames that can efficiently represent and summarize video contents and be reused in many video retrieval-related applications. An effective set of key frames, viewed as a high-quality summary of the video, should include the major objects and events of the video, and contain little redundancy and overlapped content. In this paper, a new key frame extraction method is presented, which not only is based on the traditional idea of clustering in the feature extraction phase but also effectively reduces redundant frames using the integration of local and global information in videos. Experimental results on the TRECVid 2007 test video dataset have demonstrated the effectiveness of our proposed key frame extraction method in terms of the compression rate and retrieval precision.

Keywords: Key frame extraction, Clustering, Information integration, SIFT

1. Introduction

Due to the popularity of family video recorders and the surge of Web 2.0, people produce video clips and upload their work online in an explosive speed. The increasing amounts of videos have made the management and integration of the information in videos an urgent and important issue in video retrieval. Video summarization and representation through key frames have been frequently adopted as an efficient method to preserve the overall contents of the video with a minimum amount of data [1][2]. The results of video summarization can be reused in many areas, including content-based video retrieval [3][4], semantic indexing [5], Copied Video Detection (CVD) [6], etc.

In earlier work on video summarization, key frames are selected by sampling video frames randomly or uniformly at certain time intervals [7]. This approach is simple and fast but neglects the video content. Therefore, it may miss representative frames and include redundant frames. To address this problem, shot-based key-frame extraction algorithms have been proposed [8], in which a video is first segmented into shots and then key frames are extracted for each shot independently. Key frame extraction techniques can be roughly categorized into sequential and cluster-based methods [9]. In sequential methods, consecutive frames are compared in a sequential way and key frames are detected depending on the similarity with either the previous frames or the previously detected key frame. Although the sequential methods consider the temporal ordering among frames, they only compute the similarity between adjacent frames and ignore the overall change trend in the shot range. On the other hand, in cluster-based methods, the frames are grouped into a finite set of clusters in the selected feature space [10], and then the key frame set is obtained by collecting the representatives of each cluster. In this method, temporal information of the frames is not considered; that is, key frames are selected regardless of the temporal order of each frame. If key frames are extracted for each shot independently and the scenery changes slowly in each shot, cluster-based methods are able to provide an understanding of the overall visual content of a video. In this way, the number of key frames in each shot is compact, capturing adequately the content variation along a video.

In this paper, a novel key frame extraction approach is proposed. Compared with other existing methods, our proposed approach has two contributions: (1) it extracts the key frame candidates (KFCs) using a simplified cluster-based algorithm which uses the maximum frame distance to avoid computing clusters' centers and save the calculation time; (2) it utilizes the integration of the global and local information in the video to filter the extracted key frame candidates, and then employs the local features to further refine key frame candidates, which helps the system get high quality key frames.

The rest of this paper is organized as follows. The proposed key frame extraction approach is presented in Section 2, while experimental metrics and results are provided in Section 3. Finally, in Section 4, conclusions and future work are drawn.

2. The proposed approach

The proposed approach consists of two main phases. One is the extraction of key frame candidates (KFCs) from videos, which includes a cluster-based method to extract a group of KFCs for each shot. In the filtering phase, the integration of the global and local information in the video range and the shot range extracted from KFCs is used to filter out those obvious redundant KFCs. After that, scale-invariant feature transform (SIFT) [11] is employed to extract the frame local information to further eliminate redundancy in the shot range. The system architecture of the proposed key frame extraction approach is illustrated in Figure 1.

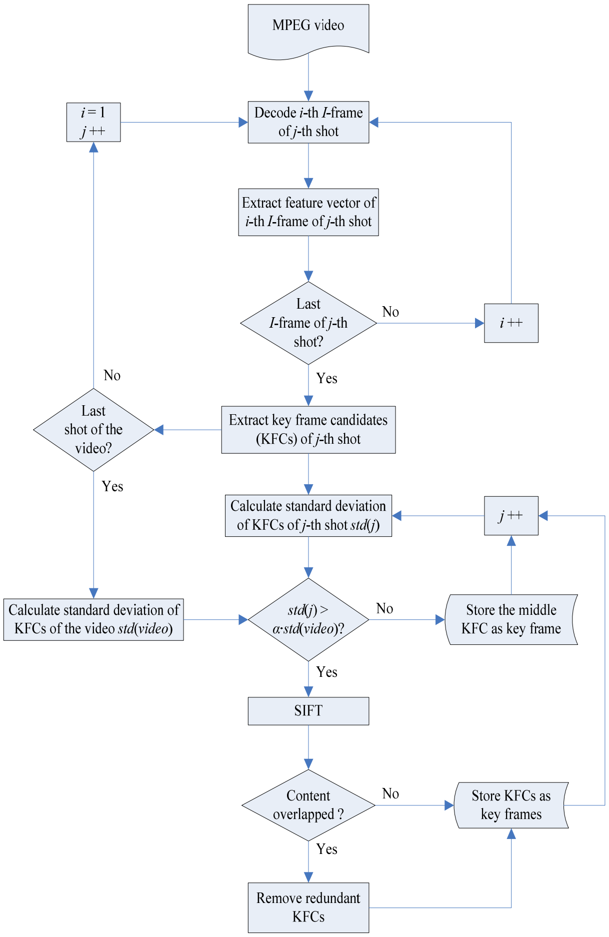

MPEG video
$i = 1$
j++
Decode i-th I-frame of j-th shot
Extract feature vector of i-th I-frame of j-th shot
Last I-frame of j-th shot?
No
i++
Yes
No
Last shot of the video?
Extract key frame candidates (KFCs) of j-th shot
Yes
Calculate standard deviation of KFCs of j-th shot $std(j)$
j++
Calculate standard deviation of KFCs of the video $std(video)$
$std(j) > \alpha{\cdot}std(video)$?
No
Store the middle KFC as key frame
Yes
SIFT
Content overlapped ?
No
Store KFCs as key frames
Yes
Remove redundant KFCs

Figure 1. **System architecture of the proposed key frame extraction approach**

2.1. Feature vector extraction from frames

In order to use more efficient descriptions to represent video content, a multi-dimensional feature vector is generated for each frame by transforming the image domain to the feature domain. In this paper, each RGB color frame is first converted into grayscale or YCbCr color space. Then the transformed frame is divided into 16×16 blocks of pixels (as proposed by Kim et al. [6]), and each block is represented by one average gray or color value. In the grayscale frame, the average gray value is the mean value of the corresponding block. For YCbCr color frame, the average YCbCr values for each block are calculated as follows [12].

$$YCbCr_{avg} = \frac{2}{3} \cdot Y + \frac{1}{6} \cdot Cb + \frac{1}{6} \cdot Cr$$

For instance, if a frame has 288×352 pixels, it will be divided into blocks and represented by a feature vector with 396 (18×22) elements. As a measure of the similarity between two frames represented by feature vectors **p** and **q**, the Euclidean distance $d(\cdot)$ is used and defined by the following equation:

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

where $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, $\mathbf{q} = (q_1, q_2, \ldots, q_n)$, and n is the number of blocks (here, n = 396).

2.2. A cluster-based method

Cluster analysis is the formal study of methods and algorithms for grouping [13]. In key frame extraction applications, the cluster-based methods take all the frames of a shot together and cluster them based on the similarity of their feature vectors. The frames of each shot form a cluster and the frame closest to the cluster's center is usually selected as the key frame. In this paper, to save the computation time, only the idea of clustering is utilized rather than actually calculating the cluster's center and its distance with nearby frames. Instead, the middle frame of each shot is used as the first KFC f_1. Based on f_1, the second KFC f_2 is chosen using the following criterion:

$$\arg\max d(f_1, f_2)$$

The n-th KFC f_n is selected by the following criterion:

$$\arg\max \sum_{k=1}^{n-1} d(f_k, f_n)$$

where k=1, 2, 3, …, n-1. This criterion shows that f_n has the largest sum of Euclidean distances with the previous n-1 KFCs.

2.3. Preliminary KFC filtering using the integrated global and local information

One major issue in key frame extraction is how many key frames should be selected per shot. Due to the unequal quantities of object information conveyed in different shots, the number of key frames to be extracted should also be different. Usually, a threshold T was employed in the key frame extractor, but the determination of threshold T is another decisive factor to affect the final performance of the key frame extractor. Instead of using statistical methods to determine the threshold's value, Chatzigiorgaki et al. [12] used two videos from TRECVid 2007 test dataset [14] as the training set to conduct the threshold selection process, which achieved good results in their experiment. However, the problem is that even though employing the training videos to calculate the threshold is acceptable, if the test video information can be utilized to decide its own number of key frames extracted in each shot, the extraction results would be more compact and accurate than those that adopt the threshold calculated based on other videos.

In this paper, the global and local frame information (e.g., standard deviation of KFCs) are used as the threshold to filter those obtained KFCs. Supposed that if the shot content changes relatively small, the value of the standard deviation of the feature vectors in the shot should be small, and vice versa. Here, the standard deviation of all KFCs is used to measure whether the content changes relatively small in the j-th shot by the following rule:

```
if $std(j) > \alpha \cdot std(video)$
    Keep all KFCs in the j-th shot
else
    Keep the middle KFCs in the j-th shot
```

where $std(j)$ denotes the standard deviation of those KFCs in the *j-th* shot, $std(video)$ denotes the standard deviation of those KFCs in the whole video, and α is the coefficient whose value is between 0 and 1. Using this rule, those obvious redundant KFCs can be filtered.

2.4. Refining key frames based on SIFT

After preliminary filtering KFCs by the global information in the video and the local information in shots, for those redundancies within shot range, more refined local information (frame range) is used by Scale-Invariant Feature Transform (SIFT) for further filtering. SIFT method transforms an image into a large collection of feature vectors, each of which is invariant to location, scale and rotation, and robust to affine transformations (changes in scale, rotation, shear, and position) and changes in illumination [11]. Based on such an ability, SIFT is considered applicable and suitable for the recognition of particular object categories in two dimensional images, three dimensional reconstruction, motion tracking and segmentation, robot localization, etc. Here, we employ the idea in SIFT to further eliminate redundant key frames with overlapped content. Assuming that if the same object is detected in two contiguous KFCs by SIFT, it means redundancy exists in the candidates and one of them should be deleted.

3. Experiments

3.1. Evaluation metrics

The evaluation of the proposed approach is carried out in terms of the percentage of the extracted key frames (compression rate) and precision-hit deviation curves. The percentage of the extracted key frames (%KF) and the compression rate are defined as follows.

$$\%\text{KF} = \frac{\text{number of extracted key frames}}{\text{total number of frames in the video}}$$

$$\text{Compression Rate} = 1 - \%\text{KF}$$

Assuming no significant scenarios were missed in the key frame extraction process, the fewer key frames we can use to represent the video, the less content redundancy existed in the key frames. Therefore, low key frame percentage is preferred.

In addition to the number of extracted key frames, the precision-hit deviation curve is introduced to evaluate the quality of the extracted key frames. We define the concept of hit deviation as the difference between ground truth frame number (target) and extracted key frame number (query).

$$\text{precision} = \frac{\text{number of } \textit{correctly} \text{ extracted key frames}}{\text{total number of extracted key frames}}$$

If two or more extracted key frames hit one target frame, the closed one was used in the calculation, while others were considered to fail to hit target. As shown in Figure 2, KF2 fails to hit adjacent ground truths GT1 and GT2, since other key frames are closer to the ground truths than it. In MPEG-1 video of 25fps, if the difference between two frame numbers is less than 25, that means the time interval between them is less than one second. In general, the frame content changes very little within one second time interval, so we consider that the two frames are similar, and key frame successfully matches the ground truth.

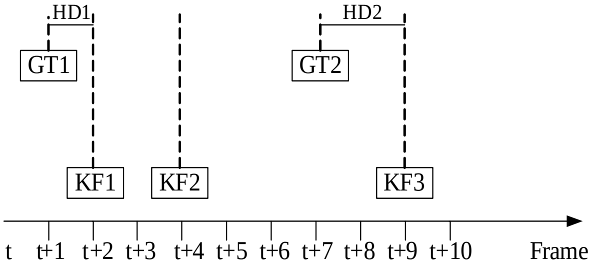


Figure 2. **Calculating hit deviation of key frames. *t* denotes video sequence at time *t;* KF1, KF2, KF3 denote key frames 1, 2, and 3; GT1 and GT2 denote ground truth 1 and 2, and HD1 and HD2 denote the hit deviation**

3.2. Testing videos

To evaluate the performance of the proposed key frame extraction approach, six MPEG-1 video sequences of 25fps from TRECVid 2007 test video collection [14] were used in the experiments. The six video's information including name, length, the number of shots, frames, and manually labeled ground truth key frames are shown in Table 1.

The grayscale and YCbCr feature vectors introduced in section 2.1 were used to extract KFCs, respectively. TRECVid provided the full annotation for the training data (using one keyframe per shot) [14]. However, since the annotation does not include the keyframe numbers, this makes it difficult to calculate the differences between the ground truth and our proposed method. We use TRECVid's annotation to manually annotate the ground truth of each test video for the comparison purpose, and adopt the result of the average sampling method as the baseline. In particular, we initially select three KFCs in each shot. In average sampling method, we averagely sample three key frames per shot. *I*-frames of each shot were first selected as basic frames for the KFC extraction.

Table1. **The six videos used to test the key frame extraction algorithms. # of KF (GT) denotes the number of key frames (Ground Truth)**

Video name	Length (hh:mm:ss)	# of shots	# of frames	# of KF (GT)
BG_2196	00:26:13	124	39339	147
BG_36511	00:10:03	72	15075	111
BG _10241	00:15:40	131	23517	131
BG_11369	00:06:33	59	9828	59
BG_37940	01:28:30	554	132759	847
BG_38002	01:08:53	700	103347	1116

3.3. Results

The experimental results of the key frame percentages and the video compression rates through key frames are presented in Figure 3 and Table 2, respectively. As can be seen from the results, the key frame percentages of our proposed approach (grayscale or YCrCb) were all limited to a maximum of 1.2 %, while on average it reached 0.8%; while in [12], the average key frame percentage is 1.5%. Comparing with the ground truths, it shows that our proposed approach effectively eliminates the redundancy and controls the key frame percentage in an acceptable range (i.e., about 1.2 times of the ground truths on average). This means that our proposed key frame extraction approach is able to extract a smaller number of key frames which are in fact more representative in summarizing the video. This is preferable in any key frame extraction method.

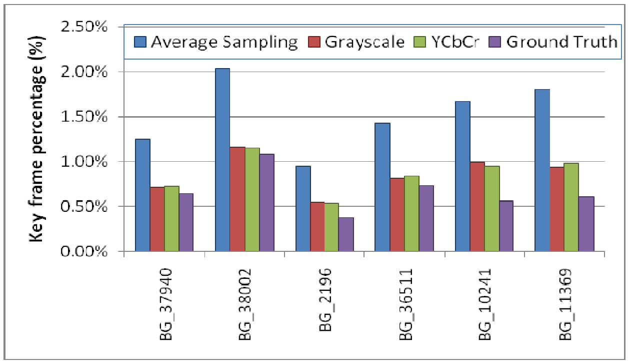


Figure 3. **Key frame percentage (%)**

Table2. **Compression rate of the six videos (%)**

Video name	Average Sampling	Grayscale	YCbCr	Ground Truth
BG_2196	99.05	99.45	99.46	99.63
BG_36511	98.57	99.19	99.16	99.26
BG_10241	98.33	99.00	99.05	99.44
BG_11369	98.20	99.06	99.01	99.40
BG_37940	98.75	99.28	99.28	99.36
BG_38002	97.97	98.83	98.84	98.92

In Figure 4, when the hit deviation was 25 frames, the precision value of our proposed approach reaches 45%, while that of the average sampling method is 35%. We also observe that the grayscale and YCrCb methods achieve similar results in Figure 4, which indicates the color factor does not play an important role at least in this series of experiments.

For qualitative evaluation, we selected the same segment of the video BG_37940 as in [12] to evaluate the proposed approach. The experimental results are shown in Figure 5. As observed from this figure, a lot of redundant frames can be seen in the baseline (average sampling) approach, even though baseline's extracting results also had little significant missing key frames. On the other hand, our proposed approach is shown to successfully reserve effective key frames and reduce the overlapped information in key frames.

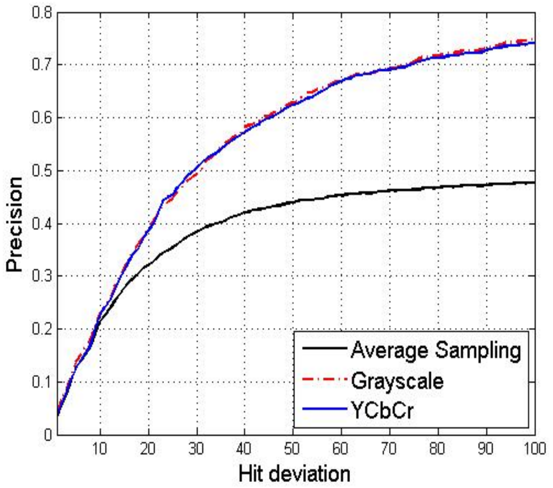


Figure 4. **Precision - hit deviation curves of the proposed approach vs. the baseline (average sampling) approach**

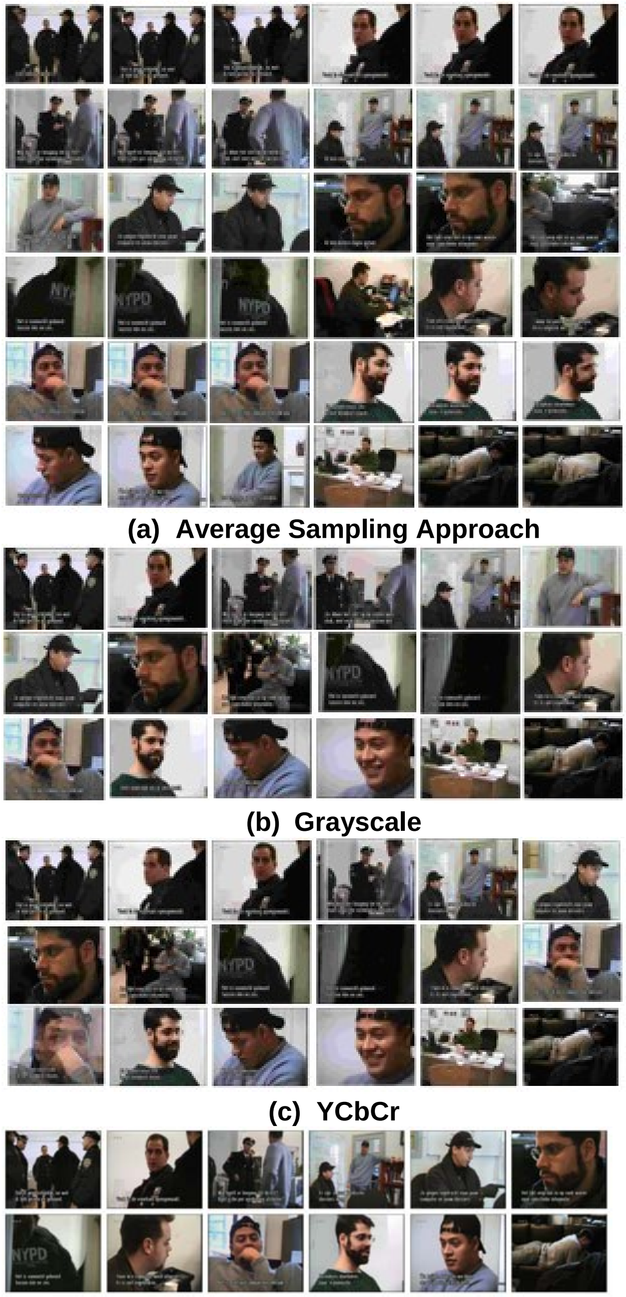
4. Conclusions and Future Work

In this paper, an effective key frame extraction approach was proposed by utilizing a clustering method to select a group of KFCs, the integrated global and local information in the video and shot ranges to do preliminary filtering, and then the KFC's local SIFT features to further refine the key frame candidates in the shot range. On the basis of the two-phase filtering, most redundant KFCs were successfully deleted, and a minimum set of key frames are obtained. The experimental results demonstrate the effectiveness of our proposed approach in terms of the compression rate and retrieval precision.

For further improvement of the proposed method, we plan to extract KFCs using object and motion information in both temporal and spatial dimensions from the video shots. We believe such information can deliver compensatory information which is not available in those static frames.

Acknowledgement

For Shu-Ching Chen, this work was supported in part by NSF HRD-0833093.



(a) Average Sampling Approach

(b) Grayscale

(c) YCbCr

(d) Ground Truth in [12]

Figure 5. **Comparison of the proposed approach with the baseline approach and ground truths on 12 consecutive shots in BG_37940**

Reference

[1] W.N. Lie and K.C. Hsu, "Video summarization based on semantic feature analysis and user preference," *In Proceedings of IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, 2008, pp. 486-491.

[2] G. Ciocca and R. Schettini, “An innovative algorithm for key frame extraction in video summarization,” *Journal of Real-Time Image Processing*, 2006, (1)1, pp. 69-88.

[3] N. Dimitrova, H. Zhang, B. Shahraray, M. Sezan, T. Huang and A. Zakhor. “Applications of video-content analysis and retrieval,” *IEEE MultiMedia*, 2002, 9(3), pp. 44-55.

[4] P. Geetha and V.Narayanan, “A survey of content-based video retrieval,” *Journal of Computer Science*, 2008, 4(6), pp. 474-486.

[5] P. Mylonas, T. Athanasiadis, M. Wallace, et al. “Semantic representation of multimedia content: Knowledge representation and semantic indexing,” *Multimedia Tools and Applications*, Springer Netherlands, September 2008, 39(3), pp. 293-327.

[6] H. Kim, J. Lee, H. Liu, and D. Lee, “Video Linkage: Group based copied video detection,” *In Proceedings of CIVR’08*, Niagara Falls, Canada, July 7–9, 2008, pp. 397-406.

[7] Y. Taniguchi, A. Akutsu, Y. Tonomura, and H. Hamada, “An intuitive and efficient access interface to real-time incoming video based on automatic indexing,” *In Proceedings of the third ACM international conference on Multimedia,* ACM, New York, USA, 1995, pp. 25-33.

[8] H. J. Zhang, C. Y. Low, S. W. Smoliar, and J. H. Wu, “Video parsing, retrieval and browsing: an integrated and content-based solution,” *In Proceedings of the third ACM International Conference on Multimedia*, ACM, New York, USA, 1995, pp. 15-24.

[9] H.C. Lee and S.D. Kim, “Iterative key frame selection in the rate-constraint environment,” *Signal Processing: Image Communication*, 2003, 18, pp. 1-15.

[10] X. Zeng, W. Hu, W. Li, X. Zhang, and B. Xu, “Key-frame extraction using dominant-set clustering,” *In Proceedings of IEEE Int’l Conf. on Multimedia & Expo (ICME’08)*, Hannover, Germany, June 2008, pp. 23-26.

[11] D. G. Lowe, “Object recognition from local scale-invariant features,” *In Proceedings of International Conference on Computer Vision*, 1999, pp. 1150-1157.

[12] M. Chatzigiorgaki and A.N. Skodras, “Real-time keyframe extraction towards video content identification,” *In Proceedings of the 16th International Conference on Digital Signal Processing*, IEEE Press, Piscataway, NJ, USA, 2009, pp. 934-939.

[13] A. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognition Letters*, vol. 31, 2010, pp. 651-666.

[14] A.F. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and TRECVid*,” In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR’06)*, Santa Barbara, California, USA, 2006, pp. 321-330.