Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 812 – 819

Procedia Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

An Improved Algorithm for Video Summarization – A Rank Based Approach

Manasa Srinivas, M. M. Manohara Pai* and Radhika M. Pai

Manipal Institute of Technology, Manipal 576 104, India

Abstract

Video summarization is one of the promising approaches for effective comprehension of video content by selecting informative frames of the video. The aim is to produce a summary of the video which is interesting to the user and representing the whole video. In this paper the proposed approach for video summarization takes various features into account such as representativeness, uniformity, static attention, temporal attention and quality which includes colorfulness, brightness, contrast, hue count, edge distribution for selecting keyframes. Experiments have been conducted on videos from open-video.org and results are compared with the standard ground truth. The results are also compared with the algorithm in literature and is found to produce better results.

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016

Keywords: Keyframes; Quality; Representativeness; Saliency; Video Summarization.

1. Introduction

The advances in storage and digital media technology has made recording and accumulation of large volumes of video very easy. Vast amount of videos are uploaded to YouTube, Dailymotion, Flickr and other video-sharing websites every minute. As hundreds of suggestions will be provided for each topic of search, browsing through these extensive videos to obtain the required video is time consuming. It is challenging to quickly retrieve this huge data efficiently. To address these challenges, efforts are being made to produce the video summary which gives gist of the entire video in short time. Video summarization is a process that facilitates faster browsing of large video collections and also more efficient in content indexing and access.

The summary can be generated either by choosing the key frames which best represent the video or through video skimming. The keyframes can be extracted using detection of change point, low level features based clustering or clustering depending on objects. The keyframes are beneficial for indexing videos but they are void of motion information. That restricts their use for certain retrieval tasks and are even less useful for enhancing the user viewing experience. In video skimming shot segments of the video will be selected for summarization. While selecting the segments, care should be taken such that it represents the whole video and also interesting to the user. However keyframes suits better for the devices with limited bandwidth and it can provide the total gist of the video in just few frames.

*Corresponding author. Tel.: +91-9945202361.
E-mail address: mmm.pai@manipal.edu

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016
doi:10.1016/j.procs.2016.06.065

The proposed technique in this paper takes quality, user attention, temporal coherence, representativeness and uniformity into account while choosing the keyframes. Video quality can be evaluated subjectively by taking user ratings or objectively on per pixel basis by considering various features like brightness, contrast, colorfulness. Visual attention based techniques extract the visually salient region which captures user attention. But both the quality and user attention considers individual frames without taking temporal information into account. In a video user will be interested in the portion where there is maximum motion, temporal coherence extracts the frames where there is maximum movement by computing inter-frame motion. The keyframes selected should represent the whole video and should be uniform and hence minimize redundant or missing data. This paper is organized as follows: In Section 2, some related works are described; the proposed approach is presented in Section 3 and finally the experimental results are discussed in Section 4.

2. Related Work

The domain specific video summarization techniques are implemented in[1,2]. However these methods cannot be used outside the domain. More universal approach for generating video summaries is discussed in[3] such as applying text processing techniques on the segments of the speech transcripts generated by automatic speech recognition. But the summary is generated solely depending on audio data by totally ignoring the visual aspects of the video. Capturing the user attention is used in[4,5] to generate the summaries. While Ejaz *et al.*[4] exploited visual, audio and linguistic cues for generating summaries, Ejaz *et al.*[5] implemented the attention curve based visual saliency detection for selecting keyframes of the video. By capturing the physiological responses of viewers, the temporal location of most salient subsegments of the video are automatically identified in[6,7]. Cong Y. *et al.*[8] expressed video summarization as a novel dictionary selection problem using sparsity consistency. Extending the idea of visual and audio curve based movie summarization as discussed in[9], the inclusion of textual cue was was proposed by Evangelopoulos *et al.*[10], so that a saliency curve derived from three methods can be used to recognize salient events which can be used to form video summaries. In Thepade *et al.*[11] discrete cosine transform coefficients of each frames are used to retrieve the frames with maximum video information. Liu *et al.*[12] proposed to exploit the low level features of image to retrieve the keyframes. The disadvantage of this method is that it ignores the high level semantic details of the video. In Thepade *et al.*[13] content based video retrieval technique is employed to retrieve the keyframes. To improve scalability and the speed of retrieval a temporal sparse approach consisting of detecting keyframes is employed. However, this method is not robust to spatial editing and hence brings the performance down. Another way of selecting keyframes is through clustering. Peng and Xiaolin[14] computed the color histogram first which was used to cluster the frames, and then the most salient frame from each cluster is selected as keyframe. The drawback of this method is temporal order of the keyframes will be lost since K-means algorithm was used for clustering.

3. Methodology

As shown in Fig. 1 keyframe selection is mainly divided into 3 stages. The scores for each frame is computed in the first stage and keyframes are selected based on the combined scores in the second stage. Finally the elimination of duplicate frames is performed in the third stage.

3.1 Quality

The quality of visual media can be affected by many factors including, but not limited to, acquisition, processing, compression, transmission, display and reproduction systems[15]. It majorly depends on the quality of the individual frames which are the building blocks of the video. The Quality score for each frame is computed using the few metrics employed in detecting the quality of frames which are described in the section 3.1.1 to 3.1.5. The average of all these individual scores for each frame is its Quality score.

3.1.1 Colorfulness

A robust and fast method to compute the colorfulness of an image using Histogram Intersection[16] is proposed. A histogram with 16 bins should be computed for each RGB channels of the frame. Similarly an Histogram is built

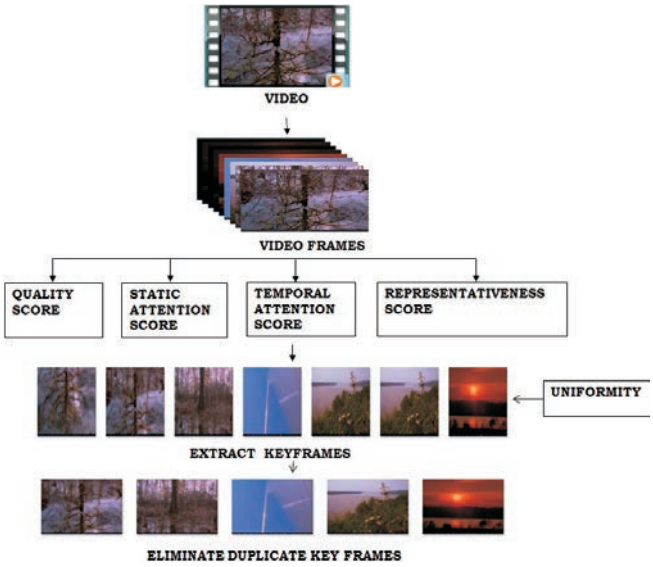
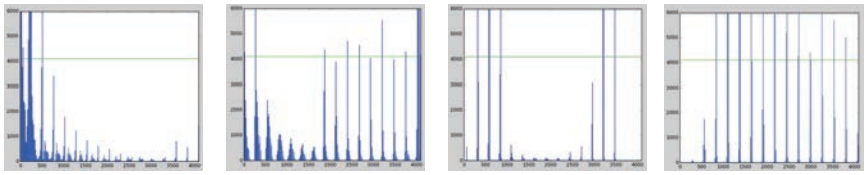


Fig. 1. Overall Framework.

(a) (b) (c) (d)

Fig. 2. Using our Colorfulness Measure, the Two Photographs (a) and (b) have High Values while (c) and (d) have Low Values.



(a) (b) (c) (d)

Fig. 3. Histogram for the Images in Fig. 2 along with the Hypothetical Image.

for an hypothetical image which is colorful i.e, contains 1 pixel of each of the possible colors in RGB space. Both the histograms are normalized to unit length and the monotonic mapping between two sets of histograms is calculated using Histogram Intersection. More the intersection value, better the similarity between the hypothetical image and probe image. The colorfulness score for images shown in Fig. 2a to d are 0.24, 0.23, 0.058, 0.023 respectively. Figure 3a to d shows the histograms where green line depicts the histogram for the hypothetical image and blue lines depicts the histogram of images in Fig. 2a to d respectively. The histogram for hypothetical image is a straight line because the hypothetical image has 1 pixel for each color and the number of pixels in each bin is constant. In case of 16 bins, the number of pixels in each bin is 4096 for the hypothetical image.

3.1.2 Brightness

Brightness is an attribute of visual perception in which a source appears to be radiating or reflecting light. Each image is converted into HSV color space and the average of value component 'V' will give us the overall brightness of the image.

3.1.3 Edge distribution

In a good quality frames, the object of interest will be well defined and hence high frequency edges will be found only at one place and not cluttered. We use the approach of[17] to compute the edge distribution.

3.1.4 Hue Count

The hue count of an image is the measure of its simplicity[17]. The number of unique hues in a high quality frame will be less though each color may be rich in tones.

3.1.5 Contrast

The human visual system is more sensitive to contrast which makes an object distinguishable. The method described in[17] is used to determine the contrast of the frame.

3.2 Static attention

Humans while watching a video will focus their attention on some specific portions or objects. While selecting the key frames it is important to identify the salient region in the frames and choose the frames with highest saliency value. The two approaches mentioned in 3.2.1 and 3.2.2 are used to detect salient region of a frame. The average score of both the methods will be the overall user attention score for each frame.

3.2.1 Region based contrast

Salient region will have high contrast values to its neighboring region than far away region. First the graph based image segmentation method is be applied on each frame and the saliency of each of the regions is obtained using the method described in[18]. More the number of pixels that belong to the salient region better the user attention score.

3.2.2 Image signature

Salient Region with image signature method is based on Discrete Cosine Transform where the sign of DCT is used in highlighting the salient region[19]. The pixels that belong to salient region will be marked as 1 and other pixels as 0. The total count of 1 will give the user attention score for the frame.

3.3 Temporal attention

In this method the changes in the values of pixels of the neighbboring frames are captured as temporal changes which is used to compute the motion information across the frames. The method described in[4] is used to compute temporal attention score.

3.4 Representativeness

The key frames we choose should be such that it represents the whole video. To select the frames which well represent the entire video k-medoids clustering[20] is used. The objective is to select the medoids such that the distance between the medoid and the data points in its cluster is minimal. The value of k depends on the number of keyframes required. This feature gives more weight to the frames which will represent the video. Initially histogram is constructed for each frame and then k-medoid clustering should be applied on this data. Representativeness score is applied to the medoid points which are the representative frames. The medoids with more data points in its cluster will have higher representativeness score.

1: For each frame 'i' compute vector of Quality(s_1), static attention(s_2), temporal attention(s_3) and representativeness(s_4) score using the methods mentioned above

$$\boldsymbol{S}_{\mathrm{i}} = [s_1, s_2, s_3, s_4] \tag{1}$$

where i=1....Number of frame(N)

2: Normalize all the scores in the range [0 , 1].Compute the standard deviation for each category of these scores.

3: The weights assigned to these scores is directly proportional to the standard deviation. The score with higher standard deviation will get maximum weightage and vice versa. The weights vector is defined as

$$\boldsymbol{W}_{\mathrm{i}} = [w_1, w_2, w_3, w_4] \tag{2}$$

4: Final score of the frames is computed as the weighted sum of each of these vectors.

$$F_{\mathrm{i}} = W_{\mathrm{i}} \cdot S_{\mathrm{i}} \tag{3}$$

where i = 1.....N and · is the dot product of 2 vectors

5: Rank the frames in the descending order of the final score where frames with top score will get higher rank. Initially choose the rank 1 frame as keyframe. Iteratively select the next frame in the order of ranking and compute the distance between the current frame and all the selected keyframes. If the distance is greater than d then add the current frame to the list of selected keyframes. Distance is computed as the difference between the frame numbers for two frames. The value of d is computed as

$$d = \frac{0.5 * N}{N_{\mathrm{k}}} \tag{4}$$

where N is total number of frames and N_k is the required number of keyframes

Algorithm 1. Primary Keyframes Extraction

3.5 Uniformity

The uniformity feature makes sure that there are no jumps while choosing keyframes. The shorter jumps might select redundant frames. Uniformity is achieved by taking temporal location of the frame i.e. its frame number. While choosing the keyframes the distance between a frame and already selected keyframes must be greater than threshold 'd'.

3.6 Proposed algorithm

3.6.1 Primary keyframes extraction

As proposed in Algorithm 1, the scores are computed for each frame using the techniques described in section 3.1 to 3.4. Each score represents different feature of a video. The weights should be assigned to each type score depending on the importance of the feature which is may vary depending on the domain and user preference. The approach followed in the proposed algorithm is to calculate the standard deviation which specifies by how much the members of a group differ from the mean value for the group, if there is more deviation then there is more deviation in the score from one frame to another. This helps us in spotting the better frames. Hence weights are directly proportional to the standard deviation. Final score of the frames is the weighted sum of each of these scores. There is a high chance that frames which are temporally close to each other will get similar score hence the top frames might belong to a same segment of the video which leads to redundant data. Hence the value of 'd' should be set such that the selected keyframes must not be too close, which also depends on the number of keyframes required by the user (N_k). Also to avoid bad frames from being selected, only top 50% frames are considered for selection.

3.6.2 Elimination of duplicate keyframes

The scenes might be of varying length in a video. Though the keyframes are selected uniformly, they might be redundant frames if the scene was long. Algorithm 2 is used to compare and eliminate the redundant frames. In this

1: Convert each key frame to gray scale.
2: Compute histogram and normalize the values in the range [0 , 1].
3: Compute the Euclidean distance between every pair of frame. If the distance is less than the threshold 't' then the frames are considered as similar and the frame with less score is eliminated from the keyframes list.

Algorithm 2. Elimination of Duplicate Keyframes

(a) Primary Extracted Keyframes

(b) After Elimination of Duplicate Keyframes

Fig. 4. Example Result.

Table 1. Dataset Details.

No.	Video Name	No. of Frames	No. of Keyframes
1	Mountain Sky water, segment 11 of 12	1160	4
2	Exotic Terrane, segment 11 of 12	3605	18
3	Challenge at Glen Canyon, segment 03 of 11	3578	10
4	Exotic Terrane, segment 10 of 12	3997	21
5	The Future of Energy Gases, segment 06 of 13	3658	20
6	Hurricanes, Segment 03	385	2
7	Hidden Fury, segment 10 of 11	1001	5

algorithm the frames are selected in the ascending order of their rank. If there are two similar frames whose distance is less than the threshold, the frame with higher score will be retained as keyframe and lower score frame will be eliminated.

4. Experiments and Results

The various set of experiments are conducted to prove the validity of the proposed framework. The testing was performed on 7 videos from open-video.org, results are compared with the ground truth keyframes provided by open-video.org. In order to estimate the results of this framework Recall, Precision and f-measure are used.

$$\text{Precision(P)} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall(R)} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

where *True Positive*(T_p): Total number of frames which are selected as keyframes and also present in ground truth, *False Negative*(F_n): Total number of frames present in ground truth but not selected as key frames by our framework. *False Positive*(F_p): Number of keyframes selected by our framework which are not present in ground truth.

$$F - \text{Measure}(F) = \frac{2 \times P \times R}{R + P}$$

F-measure is a harmonic mean of precision and recall.

Table 2. The Experimental Results by Proposed Method and Improved Frame Blocks Features Method.

	Proposed Method				Improved Frame Blocks Features Method			
No.	No. of Keyframes	P(%)	R(%)	F(%)	No. of Keyframes	P(%)	R(%)	F(%)
1	3	100	75	85.71	5	50	50	50
2	17	82.3	77.78	76.94	11	72.72	44.44	55.16
3	10	54.54	60	57.13	14	90	64.28	74.99
4	17	76.47	61.90	68.41	15	80	57.14	66.67
5	17	70.56	70	70.27	19	63.15	60	61.53
6	2	100	100	100	2	100	100	100
7	5	100	100	100	5	80	80	80

An example video is processed using the proposed framework and the results are demonstrated in Fig. 4. The initial keyframes extracted is shown in Fig. 4(a), the second and third keyframes are temporally far but they belong to the same scene which resulted in redundant keyframes. In the second stage the redundant frame is eliminated by calculating the distance between the histograms. In this experiment $t = 1.8$ gave better results.

The proposed framework is compared with the method proposed by Liu *et al.*[12], Table 1 shows the dataset details and Table 2 shows the results obtained by the proposed framework and by[12]. It can be noticed, that the key frames selected by the proposed scheme covers most of the frames from ground truth in comparison to the other technique. Further the analysis was made on the performance of the individual features on the selection of keyframes, Fig. 5 shows performance of the individual features and overall precision, recall and f-measure. Among all the features, representativeness provides better performance. It can be noticed that considering the global structure of the video while generating key frames is important.

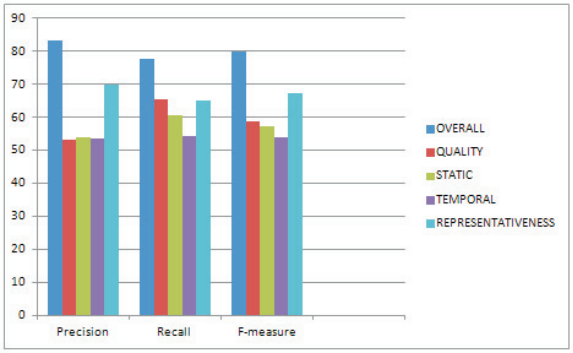


Fig. 5. Performance Comparison of Overall and Individual Features.

5. Conclusions

In this paper, an algorithm is proposed for extracting keyframes which summarizes the video. It efficiently extracts the keyframes based on the features such as quality, representativeness, uniformity, static and dynamic attention which can be used for summarizing and as well as indexing. The assignment of weights to the features based on the standard deviation allows the feature with maximum variation across the frames to get higher weights and hence helps in locating the keyframes. The experimental results obtained on the videos of open-video.org shows that the extracted key frames using the proposed scheme provides better f-measure than those generated by the other technique to which it is compared.

References

[1] F. Chen, D. Delannay and C. De Vleeschouwer, An Autonomous Framework to Produce and Distribute Personalized Team-Sport Video Summaries: A Basketball Case Study, *IEEE Transactions on Multimedia*, vol. 13(6), pp. 1381–1394, (2011).
[2] C. Xu, J. Wang, H. Lu and Y. Zhang, A Novel Framework for Semantic Annotation and Personalized Retrieval of Sports Video, *IEEE Transactions on Multimedia*, vol. 10(3), pp. 421–436, (2008).
[3] C. M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon and E. J. Delp, Automated Video Program Summarization using Speech Transcripts, *IEEE Transactions on Multimedia*, vol. 8(4), pp. 775–791, (2006).
[4] N. Ejaz, I. Mehmood and S. W. Baik, Efficient Visual Attention Based Framework for Extracting Key Frames from Videos, *Signal Processing: Image Communication*, vol. 28(1), pp. 34–44, (2013).

[5] Y. F. Ma, X. S. Hua, L. Lu and H. J. Zhan, A Generic Framework of User Attention Model and its Application in Video Summarization, *IEEE Transactions on Multimedia*, vol. 7(5), pp. 907–919, (2005).
[6] C. Chênes, G. Chanel, M. Soleymani and T. Pun, Highlight Detection in Movie Scenes Through Inter-Users, Physiological Linkage, In *Social Media Retrieval*, Springer, pp. 217–237, (2013).
[7] A. G. Money and H. Agius, Elvis: Entertainment-Led Video Summaries, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 6(3), pp. 17, (2010).
[8] Y. Cong, J. Yuan and J. Luo, Towards Scalable Summarization of Consumer Videos Via Sparse Dictionary Selection, *IEEE Transactions on Multimedia*, vol. 14(1), pp. 66–75, (2012).
[9] G. Evangelopoulos, K. Rapantzikos, A. Potamianos, P. Maragos, A. Zlatintsi and Y. Avrithis, Movie Summarization Based on Audiovisual Saliency Detection, In *15th IEEE International Conference on Image Processing (ICIP)*, pp. 2528–2531, (2008).
[10] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, Video Event Detection and Summarization using Audio, Visual and Text Saliency, In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3553–3556, (2009).
[11] S. D. Thepade and A. A. Tonge, Extraction of Key Frames from Video using Discrete Cosine Transform, In *International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, pp. 1294–1297, (2014).
[12] H. Liu and T. Li, Key Frame Extraction based on Improved Frame Blocks Features and Second Extraction, In *IEEE 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 1950–1955, (2015).
[13] S. D. Thepade and A. A. Tonge, An Optimized Key Frame Extraction for Detection of Near Duplicates in Content based Video Retrieval, In *IEEE International Conference on Communications and Signal Processing (ICCSP)*, pp. 1087–1091, (2014).
[14] J. Peng and Q. Xiao-Lin, Keyframe-Based Video Summary using Visual Attention Clues, *IEEE MultiMedia*, (2), pp. 64–73, (2009).
[15] S. Chikkerur, V. Sundaram, M. Reisslein and L. J. Karam, Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison, *IEEE Transactions on Broadcasting*, vol. 57(2), pp. 165–182, (2011).
[16] M. J. Swain and D. H. Ballard, Color Indexing, *International Journal of Computer Vision*, vol. 7(1), pp. 11–32, (1991).
[17] Y. Ke, X. Tang and F. Jing, The Design of High-Level Features for Photo Quality Assessment, In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 419–426, (2006).
[18] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr and S. Hu, Global Contrast Based Salient Region Detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37(3), pp. 569–582, (2015).
[19] X. Hou, J. Harel and C. Koch, Image Signature: Highlighting Sparse Salient Regions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34(1), pp. 194–201, (2012).
[20] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, *John Wiley & Sons*, vol. 344, (2009).