2016 Second International Conference on Cognitive Computing and Information Processing (CCIP)

Frame Clustering Technique towards Single Video Summarization

Priyamvada R Sachan
Department of ECE
East point C.O.E & Technology
Bangalore, India
priyamvada.sachan@gmail.com

Keshaveni
Department of ECE
K V G college of Engineering
Sullia, India
keshaveni@gmail.com

***Abstract*—Recent advances in technology, multimedia and social networking sites have led to a massive growth in web video content available for the general population. This results in information overload and management problem of the same. In this context, video summarization plays an important role as it aims to reduce the content size of video and yet present the important semantic concepts in the video. This gives an opportunity to reorganize video content in most succinct form for efficient and on- demand user consumption. Video summarization in its true sense is a hard problem as it involves domain specific semantic understanding of video content and user expectations. Most of the existing approaches relies on segmenting video into contiguous shots & selecting one or more keyframes from each shot and present these keyframes as summary. Such approaches may work well if independent concepts in video appear only once. However, in videos where same concepts are repeated multiple times, these existing approaches may pick repeating summary frames belonging to same concepts. In this paper, we present a novel frame clustering approach for generating very concise summaries by grouping all frames of similar concepts together irrespective of their occurrence sequence. The proposed approach is aimed towards large videos in domains like travel guide, documentaries, dramas where video revolves around few repeating concepts. The approach utilizes multiple video features in a generic way for frame-similarity determination and is extensible for multi-video summarization. Experimental comparative results substantiate the efficiency of the proposed approach in generating concise video summaries on videos with repeating concepts.**

Keywords—video processing; video summarization; similarity measure; summary frame; frame clustering; keyframe

I. INTRODUCTION

Massive increase in availability of user generated video content is already creating information flood for users. This imposes the need for enhanced video processing techniques, to enable the users to consume the desired video content effectively in a personalized manner. With excessive video content growing at large scale and constrained time with users to consume it, there emerges the need of summarization. Video summarization can be defined as a condensed representation of a video sequence conveying all important information from user's perspective from original video in the most succinct manner. More concretely, a video summary can be considered as a concatenation of a limited number of suitably chosen frames conveying maximum information from the video [1][2]. It finds various useful applications in interactive web video applications. Following section presents related work in this domain.

II. RELATED WORK

Multimedia information has shown tremendous growth of internet usage and users for last few years. Streaming video application is one such example, which has proliferated largely. In addition, significant growth has been observed in mobile and wireless devices. In such a technically booming environment, there is always a need for technologies to facilitate selection of desired and interesting video segments based on previews. This not only saves the time but also provides the user, what they desire.

Keyframe based video summarization has been in demand for many years due to its simplicity. Users can locate interesting segments by choosing a particular keyframe using a browsing tool if temporal order is not affected in selecting the keyframes. Keyframes can effectively represent the visual information of a video sequence for any web applications: video indexing, query based search engine, image retrieval techniques etc. [3]. Daniele Cerra et al. in their research work proposed, fast compression distance (FCD) to measure similarity index towards content based image retrieval [4]. The approach is sufficed to consider larger datasets for experimental purpose as compared to the existing ones, as reported in the same work.

Research in the field of video information (content) retrieval and summarization has been focusing on shot detection and clustering algorithm to extract the informative keyframes from the corresponding video frames. The authors in [5] proposed, Content-based Video Retrieval technique using keyframe extraction. An improved histogram and clustering methods for keyframe extraction are used, which are applied for shot segmentation and then from each shot, keyframes are determined based on entropy of the video frames.

[6] Presents video summarization based on color features. Here the authors have used color features with global threshold. The image difference is obtained by applying a suitable distance metric over color histogram of the images. Then by thresholding, the image difference keyframes are selected in view of video summary.

Similarity and dissimilarity-based approaches have also gained increasing popularity towards video segmentation and summarization techniques [7]. Results obtained from such approaches are more effective and accurate as compared to some of the existing methods for instance, [8].

978-1-5090-1025-7/16/$31.00 ©2016 IEEE

Video similarity measurement is a key issue for any multimedia applications such as shot detection, scene change detection, summarization etc. Video similarity measure can be based on either visual, audio or audio-visual combined features. Combined features always generate better results as compared to single feature [9][10].

III. Proposed Methodology

A video often consists of multiple scenes centered on few macro-visual concepts. These scenes may occur in a contiguous sequence or split across time (repeating concepts). A typical example for repeating concepts in video could be a travel-guide video consisting of different time interlaced repeating scenes of beaches, mountains & hotels. Most of the state-of-the-art keyframes based summarization techniques attempt to summarize (compress) contiguous scenes (frames) only. This paper presents a novel mechanism, which effectively summarizes even time-interlaced repeating frames in a video. The proposed mechanism achieves this by creating concept-clusters of video frames, where a concept-cluster is defined as a group of all semantically close frames from video centered on a single macro visual entity. These concept-clusters in essence represent different distinct concepts present in video. There after these clusters are ranked and top k clusters are selected to generate summary of a desired length. The implementation of the proposed approach is done in OpenCV Java version [11]. Fig. 1 presents the block diagram of the proposed approach. The steps involved are as follows:

Step 1: FrameSet *Generation*

A video frame is the atomic unit of video and considered as base for similarity calculations in the proposed approach. We hereby define a Frameset as a collection of contiguous frames representing same visual concept. A frameset is not necessarily equivalent to a video shot [12], instead a video shot could be comprised of multiple framesets. To segment a video into framesets, we adopt a bottom-up approach as per the following algorithm:

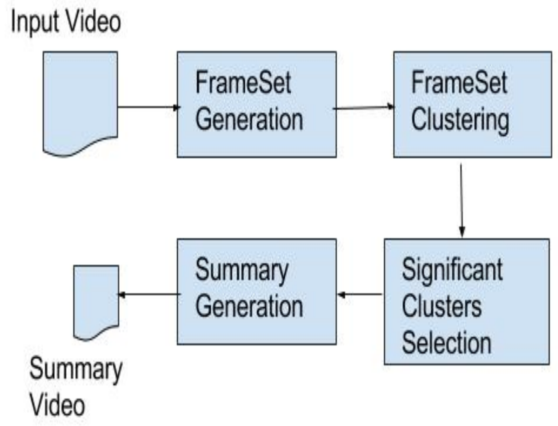

Fig. 1. Frame clustering based video summarization

FrameSet Generation Algorithm:

```
Input:
    a.  video: Input Video
    b.  sim_threshold: minimum frame similarity for considering 2
        frames as belonging to same visual concept (in our experiment
        the value is 0.85)
    c.  frameset_upperbound: maximum number of allowed frames
        in a frameset (the value is 20)
 Steps:
    1.  Decompose the input video into individual frames ($f_i$), where i
        = 1 to n; n is total number of frames in input video.
    2.  Initialize FramesetBag = {}, Fcurr= {$f_1$}
    3.  For i = 2 to n
        If size_of (Fcurr) >= frameset_upperbound
            a.  Add existing Fcurr to FramesetBag: FramesetBag
                = FramesetBag U Fcurr
            b.  Initialize new Fcurr with $f_i$: Fcurr = {fi}
       Else-If sim($f_i$, last_frame_of(Fcurr)) >sim_threshold a.
            Add $f_i$ to Fcurr: Fcurr = Fcurr U $f_i$
        Else
            a.  Add existing Fcurr to FramesetBag:
            FramesetBag = FramesetBag U Fcurr
            b.  Initialize new Fcurr with $f_i$: Fcurr = {fi}
    4.  Add final Fcurr to FramesetBag: FramesetBag = FramesetBag
        U Fcurr
    5.  Output FramesetBag as final segmentation of video
        in FrameSets
```

Input video is first divided into its individual frames. A small fixed number of time contiguous frames can be grouped together to form FrameSets. A new frame is added to growing frameset only if 1) its similarity as per frame-similarity metric with previously added frame is greater than a threshold, and 2) number of frames in frameset is less than a defined fixed upper bound; else, the new frame starts new frameset. Since a FrameSet comprises of small number of contiguous similar frames, it is expected to represent single visual concept. Fig. 2 shows the frameset generation from frames.

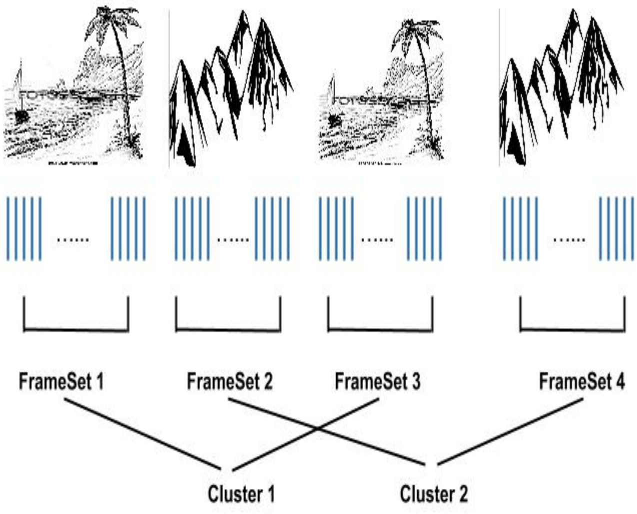

Fig. 2. FrameSet generation and clustering

Step 2: FrameSet Clustering

Next step is to group together all semantically close framesets from FramesetBag into concept-clusters such that for each distinct macro-visual concept there is a single concept-cluster as shown in Fig. 2. The framesets 1 and 3 represent the same visual concept - beach, hence they should go to same cluster - cluster 1. Similarly, framesets 2 and 4 representing mountain go to cluster 2. FrameSets are grouped using agglomerative (bottom-up) clustering technique based on following pairwise similarity function on two FrameSets:

$$Sim_{FS}\ (fs_a,\ fs_b) = \min_{(i=\{fsa\},\ j=\{fsb\})} Sim_F\ (fi,\ fj) \qquad (1)$$

Where, Sim_{FS} represents the similarity score between two framesets: frameset a (fs_a) and frameset b (fs_b). Sim_F represents the frame similarity function between two frames: frame i (fi) and frame j (fj). The above FrameSet similarity function (Sim_{FS}) relies on underlying pairwise similarity function on two frames (Sim_F). The Sim_F is defined as Euclidean distance between different features of frames:

- Color Histogram (CH)
- Edges (EI)
- Combined features

Step 3: Significant Cluster Selection

In this step, we rank the created concept-clusters in an attempt to select top k clusters as significant clusters representing top distinct concepts present in video. Here, we used the cluster size (number of FrameSets) as a basic metric for ranking the clusters. It essentially gives preference to most-seen concepts in the video. This metric was chosen primarily because it is simple to calculate and is generically applicable to all genres of video. The parameter k is configurable and largely governed by length of summary desired.

Step 4: Summary Generation

This step involves selection of frames from significant clusters towards final summary and ordering them based on their original time sequence. The desired length of summary controls number of frames to be selected from each cluster. As a default measure, we chose to select a maximum of 3 framesets from each selected cluster each contributing 1 (first) frame towards summary.

IV. Frame-Similarity Measures

Frame similarity measures play very important role towards performance of frameset generation and clustering steps in the proposed approach. Frame similarity measure can be expressed in terms of semantic distance between two consecutive frames of a corresponding video. Different distance metrics [13] can be applied for the purpose. In our experiments, we used Euclidean distance (L2-norm) that is described in next section.

A. Euclidean Distance

This is the most common distance metric in image processing and is commonly used for similarity measurement in image retrieval and indexing because of its potency. It measures the distance between two feature-vectors of images. Let Q_i and D_i be the two feature vectors then the Euclidean distance between them is calculated as given by equation 2.

$$\text{Euclidean-distance } (\Delta d) = \sqrt{\sum_{i=1}^{n} |Q_i - D_i|^2} \qquad (2)$$

Where, n represents the number of feature vectors. We applied Euclidean distance over multiple image features as described in the following section:

B. Image Features

Image features represent different visual characteristics of an image like color, shape, edge, texture etc. In our approach, we chose color, edge and a weighted combination of both color & edge as the candidates for feature vector involved in Euclidean distance for frame similarity.

1) Color Histogram *(CH):* Color is amongst one of the widely used visual image features in image /video processing research areas. Color histogram represents the distribution of different color composition in the image [14]. It focuses on the proportion of different types of colors in an image regardless of the spatial location of the colors. Although the histogram can be built for any color space, mainly it focuses on 3-dimensional color spaces like: RGB and HSV.

2) Edge Detection (EI): Edge of an image represents the sharp discontinuities, where the image shows some abrupt or gradual transform. Edge detection is a process of identifying and locating such sharp discontinuities in an image. It eliminates the extra information of an image but preserves the structural properties. The different types of edge detection operators can be used depending upon the requirement of the input and output [15][16].

3) Combined Features: Although color and edge perform individually well, both of them show some performance lag in terms of number of clusters as compared to golden data (see table 1). In view of this, we consider the blend of these features to generate combined feature set for the betterment of result. Individual features are combined by a mathematical expression given by equation 3.

$$\text{Combined feature} = x * CH + (1-x) * EI \qquad (3)$$

Where, x represents the weightage of histogram and is chosen as 0.7 for better performance. Euclidean distance metric is applied on individual and combined features as well to check the performance of the proposed approach. Table 1 presents the comparative results.

V. EXPERIMENTS AND RESULTS

We tested the proposed summarization mechanism on a dataset of two videos - a travel guide video, containing scenes from beach area, mountains and surrounding hotels, and an animal zoo video featuring activities of different animals. Both of these videos had many non-contiguous repeating concepts throughout their length. For evaluation purposes, we manually tagged the scenes and counted the number of independent macro concepts present in the videos. Main concepts observed in travel-guide video were 'beach', ‘mountains’, 'hotel', and in zoo video, were 'entry-gate', 'elephant', 'chimpanzee', 'pond', etc. For frame-similarity measure, we considered Euclidean distance over individual and combined features as mentioned in proposed methodology.

In all summarization experiments with variations on similarity measures, all concept clusters were considered with each contributing one (first) frame towards summary. The top five summary frames generated from travel-guide video using different image features are shown in Fig. 3. For evaluating summary performance in terms of concepts captured, we defined precision metric as equation 4:

$$\text{Precision} = p/\ (p+q) \qquad (4)$$

p: number of independent concepts among captured concepts
q: number of duplicate concepts among captured concepts

Same input videos were used for summary generation with an existing summarization technique based on compressive sensing clustering [17] and comparative results are shown in Table I. We observed the proposed approach with all similarity metrics performed better in identifying independent concepts accurately as reflected by precision numbers. This can be attributed to focus on non-contiguous clustering in the proposed approach, which helped in effectively grouping together frames belonging to repeating but same concepts. Existing approach attempts to group contiguous frames only and hence is not able to recognize similarity between repeating concept frames.

We also observed that frame-similarity metric indeed impacts the capability to create correct concept clusters and eventually quality of final summary. Among the three features used, combined feature over color and edge performed the best signaling that use of more image (frame) features leads to better similarity metric.

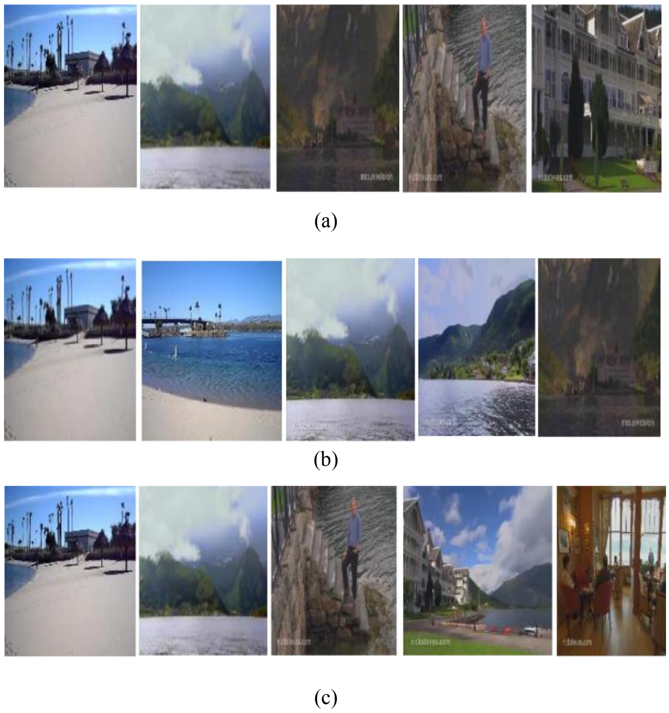

Fig. 3. Top five summary frames from travel-guide video: (a) using color feature; (b) using edge feature; (c) using combined features

VI. CONCLUSIONS

The paper presents a novel frame clustering based approach for single video summarization utilizing different visual features targeted towards creating concise semantic summaries of large videos specially in domains where distinct concepts can be repeated multiple times like travel guides, forest documentary, zoo, etc. The highlight of the proposal is creation of full video scope concept-clusters using bottom-up agglomerative clustering. The proposed mechanism relies heavily on frame-similarity metrics for effective concept clustering. The results show that, although similarity metrics based on both color histogram and edges are independently able to generate good summaries, defining a metric on combination of color & edge boosts summary quality. The base frame similarity function in the proposed algorithm is parameterized where different similarity metrics can be plugged-in. This gives flexibility to apply different similarity metrics for different domains. Since, the approach is based on frame segmentation and bottom up clustering of frames into concept clusters, it can be easily extended to topic based multi-video summarization where FrameSet clusters created from aggregate frames of multiple videos can represent common concepts across these videos.

VII. FUTURE WORK

The approach presented in the paper was tried within the scope of single video summarization with similarity metric defined on primary visual features - color and edges. The base similarity metric can be extended to include other advanced metrics utilizing multiple audio-visual features and textual subtitles. Also, the approach can be effectively extended for summarizing multiple videos around a focused topic like event news videos from multiple channels. We also plan to analyze applicability of different combinations of features as similarity metric to different domains for effective summarization.

TABLE I. Comparative Results Of The Proposed Approach

Input	Real Independent Concepts	No. of Frames	Frame SimMetric	No. of Independent Concepts	No. of Duplicate Concepts	No. of final Summary Frames	Precision
travel_guide. avi (88 secs)	24	2200	Color Histogram (CH)	14	2	16	0.875
			Edge Intensity (EI)	18	7	25	0.72
			0.7*CH+ 0.3*EI	21	2	23	0.91
			Existing method (contiguous compressive clustering)	22	11	33	0.67
zoo.avi (79 secs)	30	1975	Color Histogram	19	4	23	0.83
			Edge Intensity	23	9	32	0.71
			0.7*CH +0.3*EI	28	3	31	0.90
			Existing method (contiguous compressive clustering)	27	15	42	0.64

References

[1] Ejaz, Naveed, Tayyab Bin Tariq, and Sung WookBaik. "Adaptive keyframe extraction for video summarization using an aggregation mechanism." Journal of Visual Communication and Image Representation 23.7, 2012, pp. 1031-1040,

[2] Sujatha, C., and Uma Mudenagudi. "A study on keyframe extraction methods for video summary." Computational Intelligence and Communication Networks (CICN), International Conference on. IEEE, 2011.

[3] Mukherjee, Pratik. Shot Level Key Frame Detection using Cluster Validity Index. Diss. faculty of engineering and technology, Jadavpur University, 2013.

[4] Cerra, Daniele, and Mihai Datcu. "A fast compression-based similarity measure with applications to content-based image retrieval." Journal of Visual Communication and Image Representation 23.2, 2012, pp. 293-302.

[5] Qu, Zhong, et al. "An improved keyframe extraction method based on HSV color space." Journal of Software 8.7, 2013, pp. 1751-1758.

[6] http://www.mckvie.edu.in/site/assets/files/1389/it-4.pdf,Volume-1, Special Issue-1, 2014.

[7] http://www.ijicic.org/ijicic-13-01044. Pdf

[8] J. Jiang, X.-P. Zhang and A. C. Loui, A new video similarity measure based on video time density function and dynamic programming, Proc. of ICASSP, Prague, Czech Republic, 2011.

[9] Beecks, Christian, Merih Seran Uysal, and Thomas Seidl. "A comparative study of similarity measures for content-based multimedia retrieval." Multimedia and Expo (ICME), IEEE International Conference on. IEEE, 2010.

[10] Chang, Shih-Fu, R. Manmatha, and T-S. Chua. "Acoustics, Speech, and Signal Processing." Proceedings (ICASSP'05). IEEE International Conference on. Vol. 5. IEEE, 2005.

[11] http://docs.opencv.org/doc/tutorials/introduction/desktop_java/java_dev_intro.html.

[12] https://en.wikipedia.org/wiki/Shot_transition_detection

[13] Malik, Fazal, and BaharumBaharudin. "Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain." Journal of King Saud University-Computer and Information Sciences 25.2, 2013, pp. 207-218.

[14] https://en.wikipedia.org/wiki/Color_histogram

[15] http://dasl.mem.drexel.edu/alumni/bGreen/www.pages.drexel.edu/_we g 22/edge.html

[16] Shrivakshan, G. T., and C. Chandrasekar. "A comparison of various edge detection techniques used in image processing." IJCSI International Journal of Computer Science Issues 9.5, 2012, pp. 272-276.

[17] Pan, Lei, Xin Shu, and Ming Zhang. "A Keyframe Extraction Algorithm Based on Clustering and Compressive Sensing", 2015.