Iterative Image Based Video Summarization by Node Segmentation

Nalini Vasudevan [*] Arjun Jain [†] Himanshu Agrawal [‡]

Abstract

In this paper, we propose a simple video summarization system based on removal of similar frames and the maintenance of unique frames. It tries to capture the temporal content of the frame and to output the video with a length specified by the user. It aims at eliminating similar frames by a process of clustering where similar frames are clustered into one group. Similar frames have less degree of variation in visual frames, color distribution and visual attributes. When clusters are formed, a fraction of the frames from each of the group is retrieved to form a sequence of frames resulting in the desired output.

Keywords: video sampling, histogram, feature-frame matrix, euclidean distance, video composition

1 Introduction

Multimedia is facing a number of interesting challenges, many of them are illuminated in [2] and [7]. Video summarization is one such challenge. Automatic video content summarization has drawn attention due to its commercial potential in a number of applications. A concise video summary, intuitively, should highlight the video content and contain little redundancy while preserving the balanced coverage of the original video. Summaries are immensely useful things, especially if well made, accurate and intrinsically interesting. In our information rich age, there are many situations in which it would be useful to be able to have a single, easily understandable summary of a collection.

The growing availability of multimedia data such as video on personal computers and home equipment creates a strong requirement for efficient tools to manipulate this type of data, and to produce it in a highly efficient way to reduce time and space. Summarization can be done manually or can be automated. Automatic summarization is one of such tools, which automatically creates a short version or subset of keyframes which contains as much information as possible as the original video. Manual intervention is usually tedious and not preferred unless high intelligent video is required summaries are important because they can provide rapidly users with some information about the content of a large video or set of videos. Automatic summarization is subject to very active research, and several approaches have been proposed to define and identify the most important content in

[*] R.V. College of Engineering, Computer Science & Engineering, Bangalore, India 560059 Tel: +91 94481 07482 Email: naliniv@gmail.com

[†] R.V. College of Engineering, Computer Science & Engineering, Bangalore, India 560059 Tel: +91 99451 24241 Email: arjunjain@gmail.com

[‡] R.V. College of Engineering, Electrical and Electronics Engineering, Bangalore, India 560059 Tel: +91 98862 19916 Email: himanshu.rvce@gmail.com

a video. However, most approaches currently have the limitation that evaluation is difficult, so that it is hard to judge the quality of a summary.

In this paper, we propose a new approach for the automatic creation of summaries based on the simulated user principle, to address this problem.

2 Related Work

As quoted in [1],the goal of video summarization is to process video sequences that contain high redundancy and make them more exciting, interesting, valuable, and useful for users. The properties of a video summary depend on the application domain, the characteristics of the sequences to be summarized, and the purpose of the summary. A lot of research related activities can be found in [8] and [9]. Generally, summarization techniques try to eliminate redundant or similar frames i.e. they retain a key frame, which represents a set of frames similar to the key frame. However the continuity of the video is lost on the elimination of redundant frames, and the output result appears like a fast-forwarded video. A curve simplification strategy is employed in [6] for video summarization. Each frame is mapped into a vector of high dimensional features, and segments the feature curve into units. A video summary is extracted according to the relationship among them. [10] applies different sampling rates on videos to summarize it. The sample rate is controlled by the motion information in the shot of the video at various levels.

Video Summarization has a number of applications. Generally, depending on the application, a suitable method is chosen to summarize the video. Many such methods are discussed in [11], [12] and [13].

Video summarization is mainly achieved by selecting a set of keyframes, which vary considerably from one another.[3] uses R-Sequence method to obtain keyframes from the sampled video. However video summarization using keyframes poses some limitations. A video signal is a continuous recording of event, a spatio-temporal representation of a real time event. The distinction between images and video has to be clear. A video is not merely a set of images, but a set of images spatially and temporally related with each other. On key frame detection, the temporal characterestics of the video is lost, therefore resulting in less continuity. A set of static keyframes by no means captures these essential video properties, and is indeed a poor representation of general visual content of a video program.

Smith et al. [4] have also proposed a method of selecting key frames.The keyframes are selected based on a ranking system where a rank is given to each image, for e.g. faces and texts are given high ranks. They follow the same for the audio. Keywords are given higher preference. However, their technique requires manual intervention to rank the entire system.

In our approach, we eliminate similar frames while maintaining the continuity of the video. This method eliminates the frames using a simple division method, which is done iteratively. The boundary where the video changes are detected and frame elimination is done. It tries to preserve both the spatial and temporal locality of the input video.

3 System Overview

The entire process is shown in the figure. The first step is to sample the video and is discussed in section 4. From the images obtained, we extract the features. The method of obtaining feature frame matrix is explained

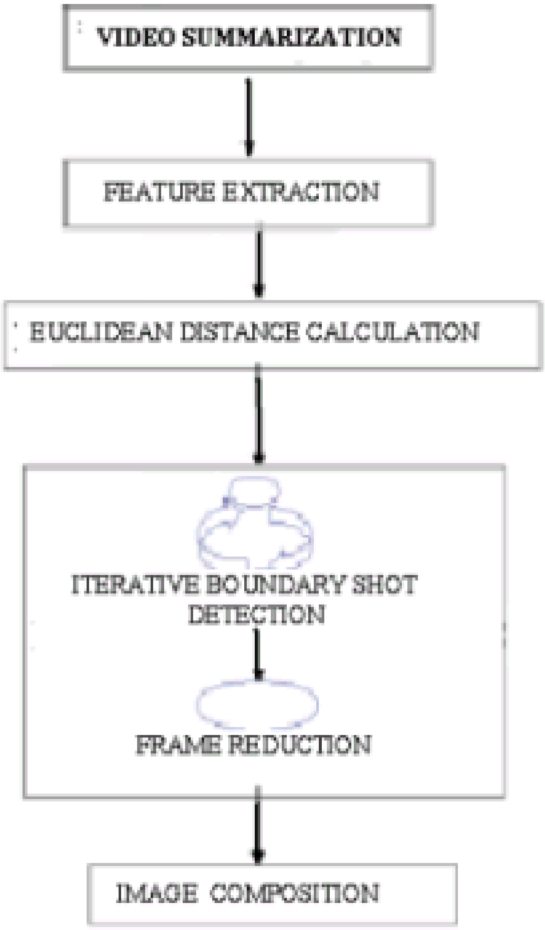

Figure 1: Process Flowchart

in section 5. We then compare the images using Euclidean distance defined in section 6. The iterative division method is applied and the steps are illuminated in section 7. The final frames are combined to form the summarized video, which is discussed in section 8.

4 Video frame extraction

The video is sampled at a constant rate. We do not consider the audio of the video. Importance is given only to the visual content. The result is a set of frames or images, which may be in gif, jpeg etc format. These images are used for the next step of the algorithm.

5 Creation of Feature Matrix

The resulting frames are analyzed to obtain the feature-frame matrix.Histograms are used because they provide a convenient way for detecting overall differences in images, and they reduce the complexities in the further steps and hence making the algorithm time-cost effective. Thus the iteration in the next step takes lesser time because the entire image is scaled down to a histogram. We use the method in [5] for constructing the histogram. Every image has RBG associated with each of its pixel. Every R or B or G is scaled down to a range 0-5 i.e. every color is allocated five bins. For R, B and G together we get 5*5*5= 125 bins. In other words, the R, B and G values of each pixel are scaled down to 0-5 and the appropriate matching bin count is incremented. To incorporate spatial information of the color distribution, each frame into 3*3 blocks, and creates a 3D-histogram for each of the blocks. These nine histograms are then concatenated together. As a result we obtain 9*125 = 1125-dimensional feature vector for the frame. This vector represents a single frame. The whole video is represented by a set of vectors, i.e. a matrix of n*1125 dimension, where n represents the number of frames in the video. Thus, the ith row represents the image characteristics of the ith frame in the video. This method is memory efficient and faster.

6 Euclidean Distance

Euclidean distance is the most common use of distance. By using this formula as distance, Euclidean space becomes a metric space (even a Hilbert space).In mathematical terms, euclidean distance measures the root of square differences between coordinates of a pair of object. The figure shows Euclidean distance D of two vectors x and y.

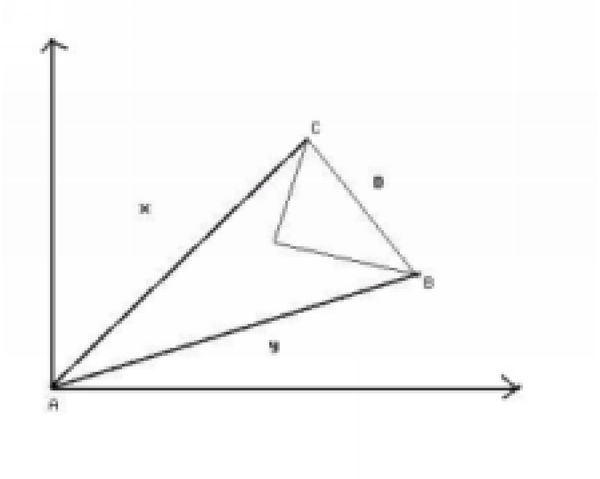

Figure 2: Euclidean distance in vector space

Once the feature frame matrix has been extracted, the boundaries where the video changes considerably have to be detected. Two frames are said to be similar when the Euclidian distance between the two frames is very less, ie less than e, where e is the error tolerance. The Euclidean distance is defined by the following: Euclidean distance is given by:

$$E = \sqrt{\sum (x_j - y_j)^2} \; for \; all \; j \qquad (1)$$

Where x and y represent 2 different images. In the feature frame matrix, it represents 2 different columns and j varies from 1 to 1125 in our case. The successive rows of the feature matrix are compared to obtain a Euclidian matrix i.e. the Euclidian distance of consecutive images is calculated. If the number of images in the original video is n, then the Euclidian matrix holds n-1 number of elements.

7 Boundary detection and node segmentation

The Euclidean distance is the primary parameter to detect boundaries. We use the following steps to detect boundaries.

1. Set depth=1. Set the number of nodes to 1; and node has values 1:n where n= number of images.

2. For all nodes do steps a-c at a given level:

   (a) Find the minimum (edmin) and maximum (edmax) Euclidian distance in the node.
   Let e_i represent the Euclidean distance between i and i+1 image. Then edmin and edmax is defined by the following.

   $$edmin = min(e_i) \qquad (2)$$

   $$edmax = max(e_i) \qquad (3)$$

   where i values represent the set of successive images of a node.

   (b) Find the approximate average Euclidian distance.

   $$edavg = (edmin + edmax)/2 \qquad (4)$$

   (c) The node is split in such a way that, if the Euclidean distance between two frames in the node is greater than edavg, a partition is drawn at that point. Partititons are drawn wherever the Euclidean distance exceeds edavg. Thus the node is split into m number of nodes along m-1 partitions, where the value of m depends on edavg.

3. depth is increment by 1

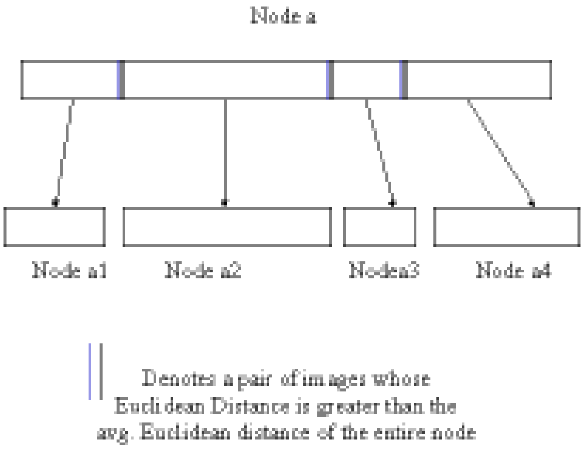
Node a
Node a1
Node a2
Nodea3
Node a4
Denotes a pair of images whose Euclidean Distance is greater than the avg. Euclidean distance of the entire node

Figure 3: Node Division

4. If depth is less than "D" [1] go to 2.
5. Let l_i be the length of original video and l_o the required length of the output video. Summarization ratio (sr) is defined by :

   $$sr = l_o/l_i \tag{5}$$

   Let nf be the number of frames in a (leaf) node. We calculate snf, which is number of frames after summarizing the node.

   $$snf = sr * nf \tag{6}$$

   The middle 'snf' number of frames calculated for every node are taken. The result is a set of frames for each node, which are concatenated together to form a sequence.

In step 2 we are finding the boundary shots in the particular node, and each node is split into sub nodes. Here, we try to find distinct shots of each node. We consider only edmin and edmax while taking the average because the node is split based on the most varying successive image pair. Once we obtain a set of leaf nodes in step 5, we try to summarize each node by a given ratio. Since the content of every frame in a node is visually close to one another, summarizing it (or eliminating similar frames) retains the key information.

7.1 Selection of the depth factor "D"

D is dependent on the kind of video. If the original video varies too much in content, then it should be large. If the video has very less changing content, a small "D" will suffice. The value D determines how deeply the video sequence should be analyzed. If the summarized video is required to capture the distinct frames, then "D" should be kept large. If the goal is to obtain continuity in the output video, then "D" is kept small. A medium value will give a compromising result usually preferred. Thus, "D" can be used as a controlling factor in determining the output.

8 Image Composition

From the previous step, we obtain a set of images. The images are combined at a constant rate, equal to the rate at which they were sampled in section 5 to obtain a moving picture or summarized video.

9 Conclusion

Video summarization algorithms generally concentrate on gathering information from one data stream, such as images, audio, or closed captions. Systematic gathering of information from all of These streams and fusing them to generate summaries will greatly enhance the summary quality. In this paper we have discussed a novel method of obtaining a summary of the video from images representing the video. The objective was to obtain a representative video containing the important frames without the loss of continuity. In

[1] D is Controlling or the Depth factor. It is discussed in detail in the next section

this method, very long shots visualizing the same content are shortened, while small shots depicting different visual content are retained. Consider a video having one constant scene. An algorithm that eliminates similar frames and retains one key frame, will give one frame as an output for this case. On the other hand, a video with a large number of variations will give improper results when the algorithm concentrates on continuity. Our approach tries to combine both the approaches and poses "D" as a controlling factor in determining the output video depending on the users choice.

References

[1] Cuneyt M. Taskiran and Edward J.Delp, "Video Summarization", CRC Press LLC (2005).

[2] Udo Hahn and Indejeet Mani, "The challenges of automatic Summarization", IEEE Computer (November 2000)

[3] Xinding Sun and Mohan S. Kankanhalli "Video Summarization Using R-Sequences ", Real-Time Imaging 6, 449-459 (2000).

[4] M.A.Smith and T.Kanade "Video Skimming for Quick Browsing based on Audio and Image Characterization", Carnegie Mellon University. Technical Report No. CMU-CS-95-186 (1995) .

[5] Yihong Gong and Xin Liu, "Video Summarization and Retreival using Singular Value Decomposition", NEC Laboratories of America. Multimedia Systems 9: 157-168 (2003)

[6] D.DeMenthon, V.Kobla and D.Doermann, "Video Summarization by Curve Simplification", ACM MM98 (1998)

[7] M.R.Naphade and T.S.Huang, "Multimedia understanding : Challenges in the new millennium", Proc. of IEEE International conference on Image Processing, Vancouver (September 2000).

[8] F.Dufaux, "Key frame selection to represent a video", Proc of IEEE International Conference on Image Processing. Vancouver (September 2000).

[9] A.Hanjalic, R.L Lagendijk and J.Biemond, "A new method for key frame based video content representation", in Image Databases and Multi Multimedia Search, World Scientific Singapore (1977).

[10] Nam.J and Tewfik.A, "Dynamic video summarization and visualization", Proceedings of ACM International Conference on Multimedia, Orlando, FL. (1999).

[11] Mark T. Maybury and Andrew E. Merlino, "Multimedia Summaries of Broadcast News. IEEE Intelligent Information Systems" (1997).

[12] Nuno Vasconcelos and Andrew Lippman, "Bayesian modeling of video editing and structure: Semantic features for video summarisation and browsing", IEEE Intl. Conf. on Image Processing (1998).

[13] Rainer Lienhart, Silvia Pfeiffer and Wolfgang Effelsberg, "Video abstracting. In Communications of ACM" (December 1997).