CSCI 576 Multimedia System Design

Final Project Report

Video Summarization

Submitted By:

Mayuresh Janorkar(USC ID:2063-2726-50)

Varun Nasery(USC ID:1767-4453-87)

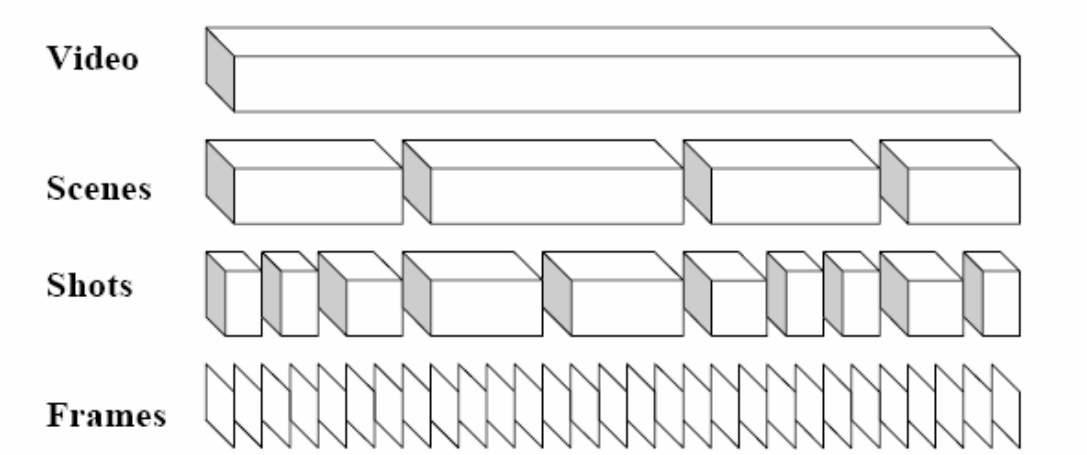
Motivation/ Abstract:

Need for video summarization-

With the advent of digital multimedia, a lot of digital content such as movies, news, television shows and sports is widely available. Also, due to the advances in digital content distribution (direct-to-home satellite reception) and digital video recorders, this digital content can be easily recorded. However, the user may NOT have sufficient time to watch the entire video (Ex. User may want to watch just the highlights of a game) or the whole of video content may not be of interest to the user(Ex. Golf game video). In such cases, the user may just want to view the summary of the video instead of watching the whole video.

Thus, the summary should be such that it should convey as much information about the occurrence of various incidents in the video. Also, the method should be very general so that it can work with the videos of a variety of genre.

Anatomy of a video-



A video nothing but a synchronous sequence of a number of frames, each frame being a 2-D image. So the basic unit in a video is a frame. The video can also be thought of as a collection of many scenes, where a scene is a collection of shots that have the same context.

In this project, we have used a single layered abstraction where a video is considered as a sequence of shots and not scenes. A shot on the other hand is a collection of frames.

Procedure:

1. Segmentation of the video into shots-

The very first step in summarizing a video is to segment the video in to multiple shots i.e. to detect shot boundaries. It is obvious that large changes in the video frame content occur at the shot boundaries. The histogram of a frame is used as a representative of the frame content. Another metric that can be used instead of a histogram is the entropy of the histogram. Entropy for a histogram is given by $E=-\Sigma\ p_i \log_2 (p_i)$, where p_i denotes the probability of occurrence of gray level 'i' in the frame.

In this project, we have used both the entropy and the histogram change as metrics for the detection of shot boundaries. If the entropy change for a frame is greater than the mean entropy change and if the ratio of histogram difference for the frame with the histogram difference of the previous frame is greater than its mean, then this frame must be a part of a new shot. For every shot, we store its shot number, starting frame, end frame, length and other metrics for the summarization criteria in a structure.

2. Computing the summarization parameters for each frame

For each frame, the parameters used for summarization- namely motion, color moments and audio level are computed.

a. Motion-

Motion in a video frame is perhaps the most important criteria used for summarization of a video. The shots that have a higher average frame motion are perceptually essential and carry more important information about what is happening in the video. Obviously, such shots with higher average motion should be included in the summary.

To compute the motion in a frame, several approaches can be employed. One of the most popular approaches is to compute the optical flow using Horn and Schunk's optical flow algorithm.

Macro-block based motion compensation can also be done for each block of a chosen size (8x8 or 16x16). In motion compensation, we determine the Euclidean distance between the macro-block in the current frame and the macro-block from the next frame which has the maximum similarity with this macro-block. However, this computation is if the order of $(2k+1)^2n^2$, where k is the size of the search window and n is the number of macro-blocks in the frame. To simplify computations, we can compute the motion vectors for only a few

macro-blocks that are chosen randomly. Care has to be taken so that each spatial region in the frame has a macro-clock chosen.

The simplest approach would be to take the difference image between the current frame and the previous frame and quantify the motion the number of pixels in the difference image that have a pixel value greater than a threshold.

b. Extraction of key-frames-

Another important criteria to decide which shots are important the number of key-frames that a shot contains. A key-frame is a frame that best represents the video content in an abstract manner. The color difference between a key-frame and its succeeding frames is not large until the next key frame arrives. On the other hand, the color difference between a key-frame and its preceding frame is quite large. Thus, a key-frame is a frame that is much different from its preceding frames and the all of its succeeding frames are similar to it.

To extract the key-frames from a shot, we compute the first three moments of color for each frame in the video. However, for ease in computation, we use the moments for only the Y channel of the image. The moments are computed using the formulae-

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} p_{ij},$$

$$\sigma_i = \left(\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - \mu_i)^2\right)^{1/2},$$

$$s_i = \left(\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - \mu_i)^3\right)^{1/3},$$

We then compute the Euclidean distance of these three moments from the corresponding moments of the previous frame. If this distance is greater than a threshold (set as equal to the mean of all such Euclidean distances), the current frame is a key frame. More the number of key-frames in a shot, more important the shot becomes in the vide summary.

c. Audio levels-

Audio also carries important information about the context of the video. For example, during an explosion in a movie or after a goal is scored in Soccer, the audio levels go high. Such scenes should have a higher chance of being a part of the video summary. Some complex procedures for analysis include analysis of various sub-bands of audio (ex. Low

frequency band, speech band, high frequency band, etc) or the processing of speech to understand the context of the video. We employ a relatively simple procedure wherein we use the average audio levels of a shot for analysis.

The frame rate of the given video is 24fps. For audio sampling rate equal to 44.1Khz, 22050 (Fs/2) samples of audio are present in every second of audio. i.e. for every video frame, 22050/24≈919 audio samples are present. These audio samples are read from the memory and their absolute values (as audio can have both positive and negative values) are summed up to give the cumulative audio level of one frame. For each shot, the average audio level is computed as the ratio of the sum of cumulative audio level of all the frames belonging to the shot to the number of frames in the shot.

d. Entropy change-

The average entropy change for a frame is also used as a metric. We compute the entropy change for each frame and the average entropy of the shot is then computed.

3. Computing the importance of each shot-

After the metrics for all the criteria have been computed, the weights for each shot are computed. All the metrics (average motion, number of key-frames, average entropy and average audio) for a shot are normalized in the range [0, 1] by dividing each metric by the maximum value of that metric for any shot. The weight assigned to each shot is the sum of each of the normalized metric. Thus, the weights assigned to the shots lie in the range [0, 4].

Ideally, a weighted sum of the three metrics should be taken where the weight assigned to each metric adaptively changes according to the video content. Ex. For a static video motion should be high, for a video with low audio levels audio should have a higher weight, etc. However, in our project, we have used the simple strategy of assigning same weight to each of the three metrics.

4. Choosing which shots should be a part of the summary-

The length of the summary video is determined by multiplying the percent value provided by the total length of the original video. After assigning weights to each of the shots, it is obvious that the shots with the highest weights should be a part of the summary video.

Also, there is a chance that the length of the most important shot may exceed the length of the summary video (when the length of summary should be very small). Each shot is assigned a ‘write bit’ which is set to 0 initially for all the shots. If the shot is chosen to be a

part of the summary video, this bit is set to 1. The following flow chart describes the logic designed to decide which shots should be a part of the summary video.

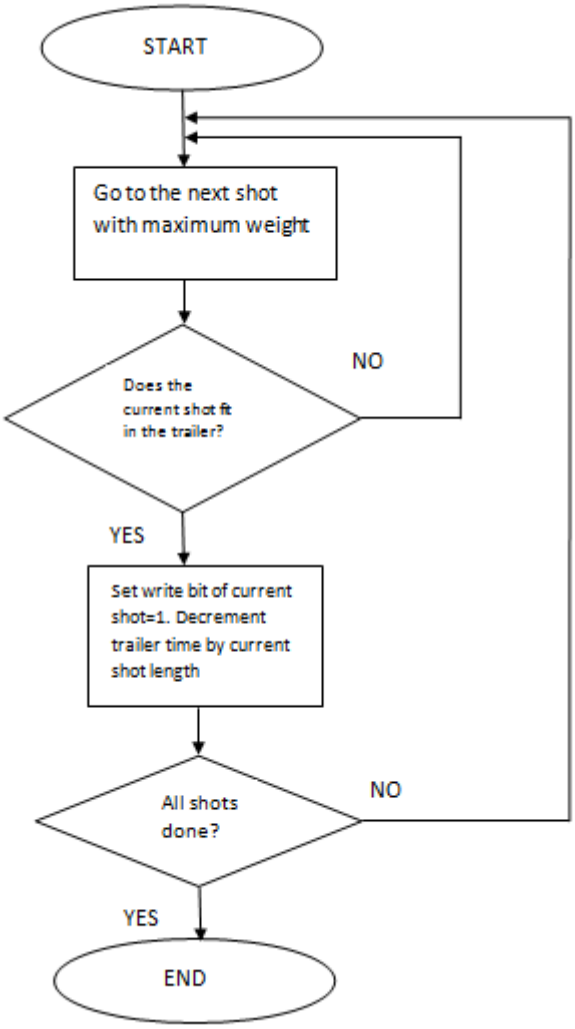


Fig. Logic describing which shots should be a part of the video

5. Writing the audio and video contents of the summary-

The shot data are scanned in a chronological order and the write bit of each shot is checked. If the bit is 1, the shot is written in the summary file. The corresponding audio data is also written in the audio summary file. After a shot is done, we proceed to the next shot. We finish the process after all the shots have been written in the summary video.

Discussion:

1. A variety of methods for scene boundary detection were tried such as histogram, entropy, both entropy and histogram. The combination of entropy and histogram used together gave best results in our case.
2. Many times, due to rapid variation in the frame content, a new shot is declared just a few frames after the beginning of a new scene. It is highly undesirable to have such small shots in our summary as it may lead to irritating discontinuities in the presentation. So, we merge shots having a length <20 frames with the previous shot.
3. The thresholds used for key frame detection, scene detection are the mean values of the metrics used for measurement. The variation in the values of these metrics is pretty high, so standard deviation cannot be used in the formula for thresholding.
4. We also tried having different weights for audio, motion and key-frames. However, most of the combinations tried did not give good results for all three videos. The combination 1:1:1 gave acceptable results for all the three videos (terminator, terminator3 and sports).
5. The summary obtained for 50% summarization was found out to be pretty comprehensive. The summary for percent value <35% sometimes seems incoherent.
6. For motion detection, two approaches were used. In the first approach we took the frame differences and quantified the motion in terms of number of pixels in the difference frame whose values are greater than a threshold. The other approach included the use of 24 macro-blocks of size 8x8. The motion vectors for these blocks were computed using a brute force search motion compensation algorithm. This approach consumes much more time and not much improvement was obtained in the results on using this method. Hence, for saving computations, this method was dropped.

References:

[1] Zhang, H.J. (1997). An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30(4), 643-658.

[2] Wolf, W. (1996). Key frame selection by motion analysis. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, GA, 1228-1231.

[3] Ngo, C., H. Zhang, and T. Pong (2001). Recent Advances in Content-based Video Analysis. *International Journal of Image and Graphics*, 2001.

[4] Kim, C., & Hwang, J. (2001). An integrated scheme for object-based video abstraction. *Proceedings of ACM Multimedia 2001*, Los Angeles, CA, 303-309.

[5] Li, B., & Sezan, I. (2002). Event detection and summarization in American football broadcast video. *Proceedings of SPIE, Storage and Retrieval for Media Databases*, 202-213.