Shot Segmentation for Content Based Video Retrieval

Nitya Raviprakash
Student, 4th Semester
Department of Computer Science and Engineering
R.V. College of Engineering
Bangalore- 560059

Meggha Suresh
Student, 4th Semester
Department of Computer Science and Engineering
R.V. College of Engineering
Bangalore- 560059

Asmitha Rathis
Student, 4th Semester
Department of Computer Science and Engineering
R.V. College of Engineering
Bangalore- 560059

Divija Devarla
Student, 4th Semester
Department of Computer Science and Engineering
R.V. College of Engineering
Bangalore- 560059

Aakanksha Yadav
Student, 4th Semester
Department of Computer Science and Engineering
R.V. College of Engineering
Bangalore- 560059

Nagaraja G.S.
Professor
Department of Computer Science and Engineering
R.V. College of Engineering
Bangalore- 560059

Abstract: Content based video retrieval is a technique used to search and browse large collections of videos stored in a database. This technique has proven to be useful for numerous applications spreadacross different domains like, surveillance, security, biomedicine, and traffic regulation. Here, the analysis is carried out based on certain properties extracted from the video frames such as colour, edge, motion, and texture. Instead of storing the features of every single frame of the video, only the features of the representative frames, which describe the entire video, are stored. This results in better storage memory utilization. We propose a method which aims at efficiently segmenting the video into shots and selecting the key frames of each shot accordingly. In order to determine the shot boundaries, we have incorporated Colour Histogram and Background Subtraction methods in this paper. The analysis of the proposed technique is carried out for different videos.

INTRODUCTION

In today's day and age, videos have become an integral part of our lives with applications ranging from multimedia to security. Consequently, obtaining the desired videos from external databases has become essential. Also, online viewing of videos has been gaining popularity. According to an analysis conducted, the number of online video users is expected to double to 1.5 billion in 2016. However, retrieving videos has proven to be quite a challenging task as the memory space required for storing the features necessary for video retrieval is quite large. Thus, reducing the memory space used is of prime importance. [1]

Presently, concept based video retrieval techniques are most widely used to obtain videos from a database. The name or description of the video in the form of text (metadata) is sent as a query for the search but this method has proven to be tedious and resource intensive because it requires a large database.[2][3]Sometimes, the keywords mentioned by the user may not properly describe the desired video. Hence, methods to carry out content based video retrieval are being researched.[4]Here,rather than sending a text-based query, the “contents” of the video such as colour, shape, edge, texture, and motions are sent as the search query.[5]

Content based video retrieval is more practical for large databases in which the videos are generated automatically, as in security surveillance cameras. The features are extracted from all the [1]surveillance videos and stored in a database as reference videos. So when a query clip is sent, its features are also obtained and is compared with the features of the reference videos. Depending on the similarities of the features, appropriate videos are sent back.[6]Video retrieval can also be used in the field of multimedia, to view videos for entertainment or educational purposes, for commercial applications like, product surveys or recognition of logos.

The process of content based retrieval involving a video-based query can by divided into four main parts; the segmentation of the videointo multiple shots, the identification of key frames torepresent the entire video, the extraction and storage of the various features of the selected frames in the database and lastly, the comparison of these features to all the other videos in the database in order to obtain related videos. [7]In this paper, the focus will be restricted to shot segmentation and determination of key frames.

RELATED WORK

Significant amount of work has been done in the field of content based video-retrieval systems. In the method proposed by T.N. Shanmugam and Priya Rajendran, the video is segmented by applying the 2-D correlation technique. For example, consider a sample video sequence comprising of a set of colour frames roughly. The first colour frame is selected from that sequence and it is transformed into grey scale format followed by the application of 2-D discrete cosine transform.[8]

According to Vegard Andre Kosmo, the usage of still images (key frames) to describe the contents of video shots is much more effective in the case of content based video retrieval. In the segment-based model, segments having their importance lower than a predefined threshold are discarded. The selected key frame of a segment will be the frame which is closest to the center of that given segment. In the shot-based approach, distinction between dominant objects is used to separate consecutive frames. Dominant objects in a frame are those colour objects with the largest number of pixels. To determine if a shot is a part of a scene, the algorithm performs a correlation score calculation. The shot is a part of the scene if the result satisfies the condition.[9]Our method combines the shot-based approach to divide the video into a number of parts as well as the segment-based model to discard unnecessary key frames.

Shweta Ghodeswar and B.B. Meshram, state that using key frames to represent scenes captures most of the content variations. Comparing each frame to every other frame in the scene and selecting the frame with the least difference from other frame requires a lot of computation and hence is not practical for most applications. Hence, it is convenient to pick the first frame as the key frame although it may not be the best match for all the frames in the scene. It will be more accurate if the frame containing the most content is selected. [10]

A Spatio-Temporal Fuzzy Hostility Index (STFHI) was proposed by H. Bhaumik et al. for determining the edges of objects present in the frames. The edges are considered as features of the frame. The correlation between the features is computed for successive frames of the video. An automatic threshold is set using the three-sigma rule, and hard cuts are detected in the video. [11]

The properties of the HSV colour space regarding visual perception of the variation in Hue, Saturation and Intensity values of an image pixel have been studied in the paper proposed by Shamik Sural et al. For image segmentation, pixel features are extracted by either choosing the Hue or the Intensity as the dominant property based on the Saturation value of a pixel. The feature extraction method has been applied histogram generation in content based image retrieval. Segmentation using this colour space shows better identification of objects in an image. The histogram retains a uniform colour transition. The results generated are better on comparison with those generated using the RGB colour space.[12] Hence, the HSV colour space is used in the Content Based Video Retrieval method proposed as well.

METHODOLOGIES

In order to minimize the amount of space required to store the information about the video, the video is first divided into a number of shots i.e. shot segmentation. A video can be segmented into shots in various ways. In the method employed here, we mark the shot boundary whenever a change in scene is detected by comparing the colour histogram of the frames. Further, background subtraction is used to obtain key frames containing the objects in motion. If greater specifications is necessary, multiple frames are selected to represent the same. This will be useful to store different angles/faces of an object within a shot which may be important.

SHOT SEGMENTATION

A shot is basically a series of consecutive frames taken by a single camera without any major change in the colour composition of the different frames comprising it.[2]As a shot boundary occurs whenever there is a distinct change in the colour contents of two frames, the shot segmentation process can be carried out by comparing the colour histograms of every pair of consecutive frames. A colour histogram gives us a representation of the colour distribution in an image. It shows the number of pixels corresponding to different colours.

Often, colour histograms are built using either the RGB colour space or the HSV colour space. HSV consists of three matrices; hue, saturation, and value. ‘Hue’ indicates the colour, ‘saturation’is the amount of grey in the colour (greyscale), and ‘value’ describes the brightness or intensity of the colour. In video processing, it is preferable to use the HSV representation because it separates the Intensity (luminance) from the colour information (chromaticity).[13] Sometimes, this property can be used to detect lighting changes or for removing shadows. A HSV colour histogram shows the number of pixels corresponding to different hues, saturations, and values depending on the number of channels considered.

In order to roughly determine the shot boundaries, we determine the coefficient of correlation between the colour histograms of two consecutive frames starting from the beginning of the video.[14] This is calculated using the formula.

$$d(H_1, H_2) = \frac{\sum_I (H_1(I) - H_1)(H_2(I) - H_2)}{\sqrt{\sum_I (H_1(I) - \bar{H}_1)^2 \sum_I (H_2(I) - \bar{H}_2)^2}}$$

$$\bar{H}_k = \frac{1}{N}$$

H_1 and H_2 represent the histograms being compared and N indicates the total number of bins.

The value of the correlation coefficient can range from -1 to +1. Higher the value of the metric, greater is the accuracy of the match. We roughly detect the shot boundaries by assuming a threshold value which lies between 0.6 and 0.8 depending on how hard the cut is. If the correlation between two consecutive frames is less than the threshold value, this would indicate that the colour distributions in the two frames are very different from each other. This means that a scene change has taken place and thus a shot boundary is detected at this point.

One of the disadvantages of assuming the threshold value is that for a soft cut, due to the gradual transition from one shot to another, numerous shots may be identified within a relatively small number of frames. However, it is not efficient to use all of these shots in video analysis as they represent almost the same information and this unnecessarily occupies more space in the memory. Therefore, we discard the redundant shots.

BACKGROUND SUBTRACTION

Once the rough shot boundary is detected, background subtraction of the frames (also known as foreground detection) is carried out. This technique is used for extracting an image’s foreground i.e. objects in motion, for further processing. Using background subtraction, a reference frame is identified in each shot to detect the moving object.[15] Each image is then subtracted from the reference image in terms of pixels.

We follow the adaptive method wherein the reference frame is constantly updated. This helps eliminate problems with frequently moving objects which otherwise would have become a part of the background. In this paper, an adaptive technique called mixture of Gaussian is specifically addressed. It is further divided into two methods: MOG, MOG2.

MOG technique uses a mixture of k Gaussian distributions (k=3 to 5 by default). It tracks multiple Gaussian distributions simultaneously. The weight of the mixture represents the time proportion for which the colour remains in the scene. Colour of the background is chosen based on its weight. Since MOG is parametric, the model parameters can be adaptively updated without keeping large buffer video frames. In MOG, the background ratio is the threshold value that signifies if an object is included in the background or not. This value is by default considered to be 0.7.

DETECTION OF REPRESENTATIVE FRAME

There are three cases that need to be considered: Shot with no distint object, shot with a single object and a shot with multiple objects. For the first case, the first frame of the shot is taken the representative frame. In the next case with a single object, the frame in which the object enters is considered as the representative frame,using background subtraction. Here, the newly obtained representative frame replaces the representative frame obtained from the previous case. Lastly, we increase the threshold value in the previously mentioned shot segmentation technique to 0.9 so as to detect all the objects in the shot, for shots with multiple objects. These extra frames are also considered as representative frames of the same shot and the corresponding frame numbers are stored in a 2-D data structure.

EXPERIMENTAL RESULTS

When we carried out the shot segmentation process, figure 1 and figure 2 were detected as the end of one shot and the beginning of the next shot respectively. Whereas, figure 2 and figure 3 are parts of the same shot although the frames are different.

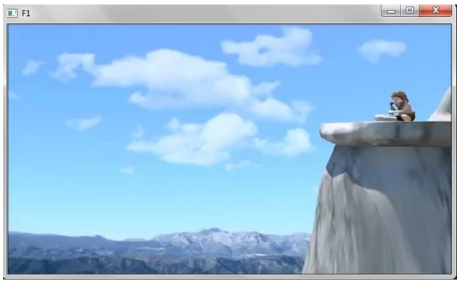


Figure 1

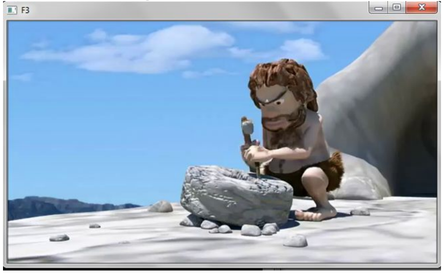


Figure 2-Hammer Up

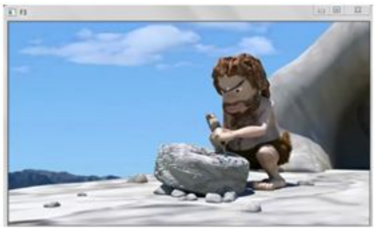


Figure 3-Hammer Down

We performed background subtraction for a shot. The first frame of the shot is shown in Figure 4. The frame that depicts the entry of the object is shown in Figure 5 which replaces the first frame as the key frame. Figure 6 shows the foreground object of Figure 5.

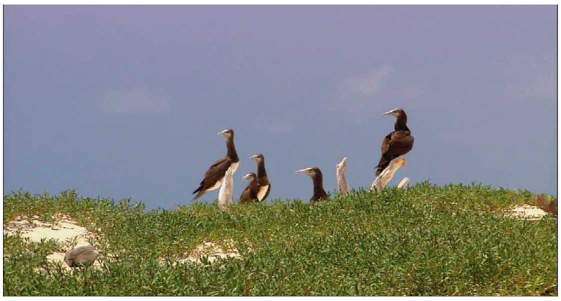


Figure 4- Frame number 126

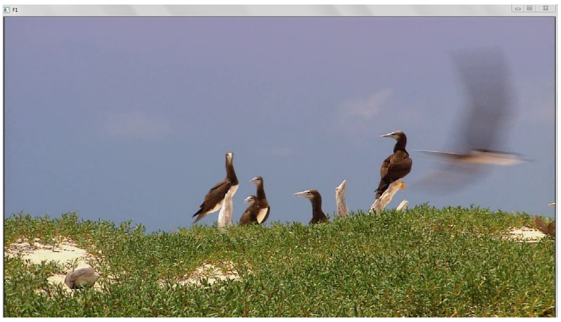


Figure 5

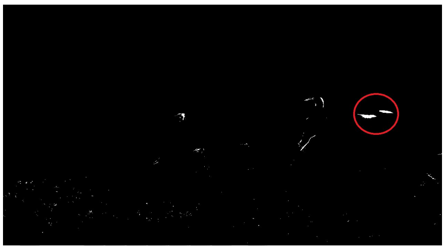
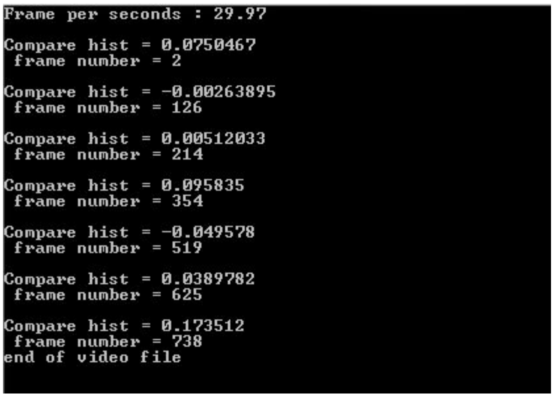


Figure 6- Background Subtraction (Bird Detected)



```
Frame per seconds : 29.97

Compare hist = 0.0750467
 frame number = 2

Compare hist = -0.00263895
 frame number = 126

Compare hist = 0.00512033
 frame number = 214

Compare hist = 0.095835
 frame number = 354

Compare hist = -0.049578
 frame number = 519

Compare hist = 0.0389782
 frame number = 625

Compare hist = 0.173512
 frame number = 738
end of video file
```

Figure 7

Figure 7 shows the frames at which the shot boundaries occur along with the correlation coefficient between that frame and the previous frame. Since the correlation values are below the threshold of 0.7, a change in scene is detected.

CONCLUSION AND FUTURE WORK

This technique aims at storing the features of the video effectively in the database as well as representing the video contents accurately. This type of representation will be especially useful in fields like medical diagnosis and satellite imagery wherein a large number of specifications are required for analysis. For example,in medical multimedia retrieval, a text-based query may lead to ambiguity. Hence, it is easier to provide the exact requirements through a content-based query. However, in order to perfectly retrieve the required videos, a combination of both content based and concept based video retrieval may be preferable. We plan to extract the features of the determined representative frames and store them. Further, the features are compared with the features of all the videos in the database in order to retrieve the related videos thus completing the video retrieval process.

REFERENCES

[1]. I. Sezan, "Applications of Analysis and Retrieval," pp. 42–55, 2002.

[2]. A. Amiri, "VIDEO SHOT BOUNDARY DETECTION USING GENERALIZED EIGENVALUE DECOMPOSITION AND GAUSSIAN TRANSITION DETECTION Ali Amiri Mahmood Fathy," vol. 30, pp. 595–619, 2011.

[3]. A. P. Natsev, I. B. M. T. J. Watson, and A. Haubold, "Semantic Concept-Based Query Expansion and Re-ranking for Multimedia Retrieval A Comparative Review and New Approaches."

[4]. N. Carolina, M. Hall, C. Hill, B. M. Wildemuth, and G. Marchionini, "The relative effectiveness of concept-based versus content-based video retrieval," pp. 2–5, 2004.

[5]. M. Petkovic, "Content-based video retrieval."

[6]. T. Horprasert, D. Harwood, and L. S. Davis, "A Statistical Approach for Real-time Robust Background Subtraction and Shadow Detection," pp. 1–19.

[7]. H. Pituach, S. Shamay, U. Lavi, H. Pituach, F. Application, and P. Data, "(12) United States Patent," vol. 1, no. 12, 2014.

[8]. P. Rajendran, "AN ENHANCED CONTENT-BASED VIDEO RETRIEVAL SYSTEM BASED ON QUERY CLIP," vol. 1, no. 3, pp. 236–253, 2009.

[9]. V. A. Kosmo, "Efficient Algorithms for Video Segmentation," no. June, 2006.

[10]. S. Ghodeswar and B. B. Meshram, "Content Based Video Retrieval 1," pp. 2–5.

[11]. H. Bhaumik, S. Bhattacharyya, and S. Chakraborty, "Video Shot Segmentation Using Spatio-temporal Fuzzy Hostility Index and Automatic Threshold," 2014 Fourth Int. Conf. Commun. Syst. Netw. Technol., pp. 501–506, Apr. 2014.

[12]. G. Q. and S. P. Shamik Sural, "SEGMENTATION AND HISTOGRAM GENERATION USING THE HSV COLOR SPACE FOR," pp. 589–592, 2002.

[13]. K. T. Rd, "GRADUAL TRANSITION DETECTION USING COLOR COHERENCE AND OTHER CRITERIA IN A VIDEO SHOT META-SEGMENTATION FRAMEWORK Efthymia Tsamoura , Vasileios Mezaris , Ioannis Kompatsiaris Informatics and Telematics Institute / Centre for Research and Technology Hellas," pp. 45–48, 2008.

[14]. J. E. L. Hyeon Jun Kim, "(12) United States Patent," US7376263 B2, 2008.

[15]. J. K. Suhr, H. G. Jung, G. Li, and J. Kim, "Mixture of Gaussians-based Background Subtraction for Bayer-Pattern Image Sequences," 2010.