Journal of Computer Science 8 (8): 1364-1371, 2012
ISSN 1549-3636

Shot Detection Using Genetic Edge Histogram and Object Based Video Retrieval Using Multiple Features

[1]R. Kanagavalli and [2]K. Duraiswamy
[1]Anna University, Chennai, India
[2]Department of Computer Science and Engineering,
KS Rangasamy College of Technology, Kalvi Nagar, Tiruchengode-637215, India

Abstract: As the usage of multimedia data increasing rapidly, how to get the video data we need efficiently become so important. Recent advances in multimedia technologies allow the capture and storage of video data with relatively inexpensive computers. **Problem Statement:** However, without appropriate search techniques all these data are hardly usable. Users want to query the content instead of the raw video data. Today research is focused on video retrieval. Content-based search and retrieval of video data becomes a challenging and important problem. To retrieve the content of the video the user need automatic classification and categorization of the visual content. **Approach:** In this study a novel algorithm is proposed for shot detection using Genetic Edge Histogram and 2-D discreate cosine transform as a feature and multiple features like color, motion, shape and SIFT are used to retrieve the similar shots. **Results and Conclusion:** The combination of proposed features yields good results interms of precision and recall.

Key words: Video retrieval, feature extraction, SIFT, shot detection

INTRODUCTION

Video has a strict hierarchy in nature. It is organized as shown in Fig. 1. Each video is divided in to scenes and each scene consists of meaning full shots. A shot is defined as one or more frames. Shot represents a continuous action in time or space. A frame has a real world objects in it.

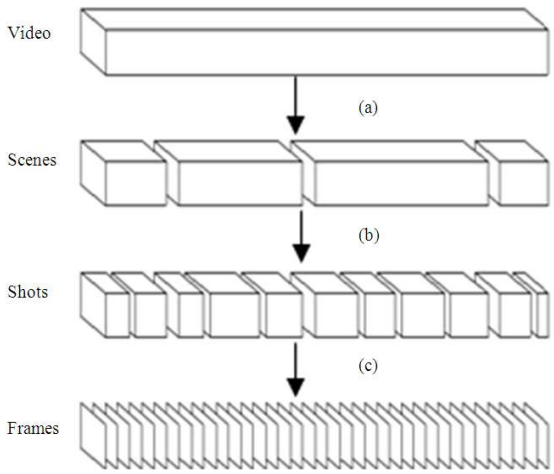


Fig. 1: (a) Video into scenes (b) Scenes into shots (c) shots into Frames

Retrieval of multimedia data is mainly divided into two parts:

- Text Based Retrieval
- Content Based Retrieval

Text retrieval is an easy process and it has been used for long time. But it has so many problems like manual text annotations, selecting a text for process. In content based retrieval spatial features like color, shape, texture and the temporal features like motion can be used to retrieve videos.

MATERIALS AND METHODS

The architecture of content-based video retrieval system: The system extracts the various visual features present in the video and then stores it in the database. Before storing the various features they are indexed. The user then gives a query to the system. The features are extracted from the query and are matched with the database and the retrieved shots are displayed (Gunsel and Tekalp, 1998: Flickner *et al*., 1995; Smith and Chang, 1996). In the proposed work the video is first divided into frames. Each frame is an image. The numbers of frames are based on the length of the video. In the proposed work the process of retrieving the video is performed as follows:

Corresponding Author: Kanagavalli, R., Anna University, Chennai, India

- The video is divided into frames
- These images are grouped into shots using Genetic edge Histogram and DCT coefficient based block matching method
- RGB to Lab conversion is performed
- Colour Quantization is performed inorder to extract colour feature
- Motion feature is extracted
- Shape feature is extracted
- SIFT feature is also obtained for all the frames in the shot
- For the query object also all the affordable features are extracted
- Comparison is performed between query object and the features in the feature library
- Similar frames are retrieved

Shot segmentation: Shot is the sequence of frames that contains the real world objects. The literature reports plenty of effort for shot segmentation (Anjulan and Canagarajah, 2006; 2007). We performed shot segmentation using edge detection and 2-D Correlation Coefficient based block matching (Thakre *et al*., 2010). First the edges are identified in two consecutive frames to make the algorithm as less sensitive to small changes.

First frame is extracted from the video and is divided into 8×8 blocks. For each block 2-D Discrete Cosine Transform is calculated. The same process is performed for the adjustant frame also. Block wise comparison is performed for the two frames. This process is performed for all the frames in the video. If the comparison exceeds the certain threshold then the shot is detected. Similar shots are grouped together. DCT is a separable linear transformation; that is, the two-dimensional transform is equivalent to a one-dimensional DCT performed along a single dimension followed by a one-dimensional DCT in the other dimension. The following equation is used to obtain DCT for each block of the image. The definition for 2 Dimensional DCT for the input block A and the output block B is Eq. 1:

$$B_{pq} = \alpha_p \alpha_q \sum_{M-0}^{M-1} \sum_{N-0}^{N-1} A_{mn} \cos\frac{\pi(2m+1)}{2M} \cos\frac{\pi(2n+1)}{2N} \quad (1)$$

Where:

$$\alpha_p = \begin{cases} \frac{1}{\sqrt{M}}, & p = 0 \\ \sqrt{\frac{2}{M}}, & 1 \le p \le M-1 \end{cases}$$

And:

$$\alpha_q = \begin{cases} \frac{1}{\sqrt{N}}, & q = 0 \\ \sqrt{\frac{2}{N}}, & 1 \le q \le N-1 \end{cases}$$

where, M is the row size and N is the column size of the block. The correlation coefficient for the comparison blocks are calculated based on the following formula Eq. 2:

$$R = \frac{\sum_m \sum_n (A_{mn} - A)(B_{mn} - B)}{(\sum_m \sum_n (A_{mn} - A)^2)(\sum_m \sum_n (B_{mn} - B)^2)} \quad (2)$$

Shot is identified if the mean of the feature value exceeds certain threshold value. Figure 2 and 3 shows the frames from the shot for the input videos.

Genetic edge histogram: The next method for finding the scene change detection is GEH (Genetic Edge Histogram). The overview of the method is described as follows. A frame is taken as input and binomial filter is applied to smooth the image. The resulting smoothed image is then applied gradient which will result in two images corresponding to horizontal and vertical components. Both the images are carried out with non-maxima suppression. From the result of nom-maxima suppression algorithm, a binary image is produced using two levels of threshold, such as lower and upper thresholds. From the generated binary image, edge histogram is computed and is considered as a scene change detection parameter for that frame.

In order to choose an optimized lower and upper threshold levels, genetic algorithm is implemented here. The following section explains how genetic algorithm is used for the threshold calculation in detail.

Genetic based optimized threshold calculation: Initially 4 parent chromosomes with 7 genes for each are chosen as the parameters for genetic optimization. Out of 4 chromosomes, 2 chromosomes are assigned for lower threshold (LC1 and LC2) and another 2 chromosomes are assigned for upper threshold (UC1 and UC2) as shown if Fig. 4. The lower threshold value is taken from 0 to 0.4 and upper threshold value is 0.5-1.0.

In the first step, LC1 and UC1 genes values are randomly selected from its corresponding threshold ranges respectively. Thus for each lower and upper threshold value from LC1 and UC1, the previously available non-maxima suppressed image is converted into binary image. The binary labeling process is carried out after conversion process. The total number of labels is calculated. So, we will get 7 values totally. These values are sorted and the middle value is taken.

Fig. 2: Sample frames from shot

Fig. 3: Sample frames from shot

G1	G2	G3	G4	G5	G6	G7

Parent chromosome for lower threshold (LC1)

G1	G2	G3	G4	G5	G6	G7

Parent chromosome for lower threshold (LC2)

G1	G2	G3	G4	G5	G6	G7

Parent chromosome for upper threshold (UC1)

G1	G2	G3	G4	G5	G6	G7

Parent chromosome for upper threshold (UC2)

Fig. 4: Initial chromosomes and genes selection

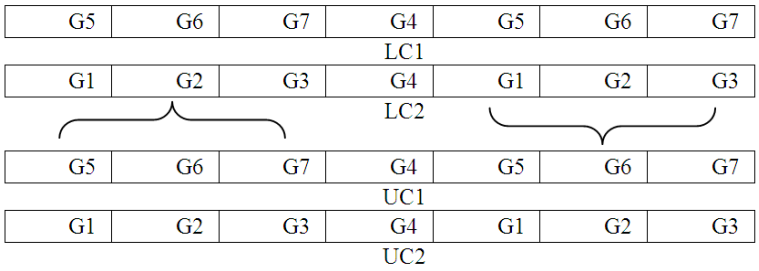


Fig. 5: Genes position after crossover

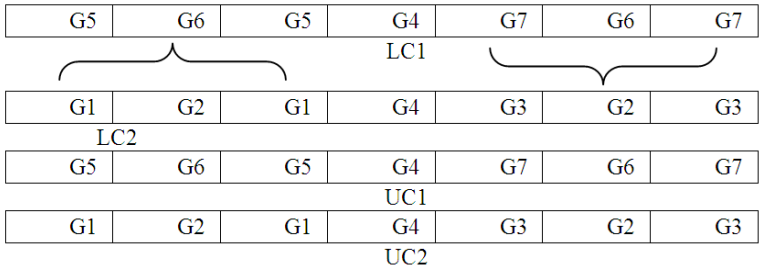


Fig. 6: Genes position after mutation

Table 1: Upper threshold, lower threshold and best fitness value for different iterations

Iteration	Upper Thrs.	Lower Thrs.	Fitness
100	2.0123	2.0241	17852
200	3.5874	2.9245	12879
300	4.4879	3.8678	8542
400	5.2615	4.9578	5097
500	5.9945	5.3874	3721
600	7.4587	5.7751	2475

Table 2: Best fitness and number of connected edges for different iterations

Iteration	Best fitness	Number of connected edge
100	17852	24587
200	12879	18584
300	8542	16254
400	5097	9125
500	3721	8956
600	2475	8454

This middle value is the fitness value (F1). The same produce is repeated for LC2 and UC2 chromosomes to get another fitness value (F2). Then single bit crossover process between (LC1, LC2) and (UC1, UC2) is carried out in which the gene values are shuffled respectively across the chromosomes. After the crossover, single bit mutation is done in which gene values are interchanged within the chromosomes. Figure 5 and 6 shows the gene position after the crossover and mutation process.

After the crossover and mutation, the above mentioned fitness value computation procedure is repeated. Thus we arrive at another two fitness values F3 (LC1, UC1) and F4 (LC2, UC2). Now, minimum of F1, F2, F3 and F4 is calculated and the gene values of corresponding chromosome pair are kept unchanged. The rest of the chromosome genes are replaced with new set of random values and above mentioned procedures are repeated for multiple iterations until a best fitness value is obtained. The chromosome pair for which best fitness value is found out is chosen for final threshold selection. Binary images are produced for each lower and upper threshold values available in the best chosen chromosome and the connected components for these images are calculated. The median value of total connected components is computed which in turns return corresponding final lower and upper threshold values respectively. The binary image obtained for using final threshold bounds is undergone edge detection algorithm and the number edges is calculated which is then used as scene change detection parameter. When the iterations of GA are increased, the connected components will vary largely in different iteration until it converges to minimal difference value. Table 1 is the fitness results obtained for various iterations and Table 2 is the computed number of edges for various iterations.

Figure 7 shows the pseudo code for Genetic edge histogram feature calculation.

Object based Video Retrieval: After the shot segmentation, object based video retrieval is performed by extracting the features like colour, motion, SIFT and shape. The required object can be segmented from the frame using segmentation process (Kim *et al*., 2008, Zhong and Chang, 1997; Jain, 1989). Object based video retrieval is the active field of research.

```
Algorithm for GEH:
Require: I, Lt, Ut
Return: Fv
I→Image, Lt→Lower threshold, Ut→Upper threshold
Ibs→ do Binomial_Filters
IGxy→ compute x and y gradient in Ibs
Inmx→ Compute Non-maximum suppression in IGxy
DoProcessGABased_Edge_Detection(Inmx,Lt,Ut)
Process GA Start
G1→get rand Lower limit threshold
G2→get rand Upper limit threshold
Process→Crossover, Mutation
  Start Process→Fitness
      Bl→Binary Label
      Fv→MID (SORT (Bl)) →Initial Selection
      WHILE Loop<100
          FOR S1
              FOR S2
                  IF Bl(S1,S2)>Lt and Bl(S1,S2)<Ut
                      Bl(S1,S2) assign 1;
                  END IF
              END FOR
          END FOR
      Bl→Binary Label
      Fv→MID (SORT(Bl)) (best selection)
      END WHILE
  End Process
Process GA End
```

Fig. 7: Algorithm for genetic edge histogram

Colour feature extraction: Color is an important feature in video retrieval because of its effectiveness and simplicity. In this study with the aid of color quantization the color features of the frames are extracted. First each frame is converted from RGB color space to Lab color space .RGB cannot be converted to LAB directly. First it should be converted to XYZ color space then to LAB as follows:

$$[X] = [0.412453\ 0.357580\ 0.180423]\ [R]$$
$$[Y] = [0.212671\ 0.715160\ 0.072169]\ [G]$$
$$[Z] = [0.019334\ 0.119193\ 0.950227]\ [B]$$

$$L^* = 116 * (Y/Yn)^{1/3} - 16 \text{ for } Y/Yn > 0.008856$$
$$L^* = 903.3 * Y/Yn \text{ otherwise}$$

$$a^* = 500 * (f(X/Xn) - f(Y/Yn))$$
$$b^* = 200 * (f(Y/Yn) - f(Z/Zn))$$

where $f(t) = t^{1/3}$ for $t > 0.008856$ $f(t) = 7.787 * t + 16/116$ otherwise.

Fig. 8: Lab colour space frames

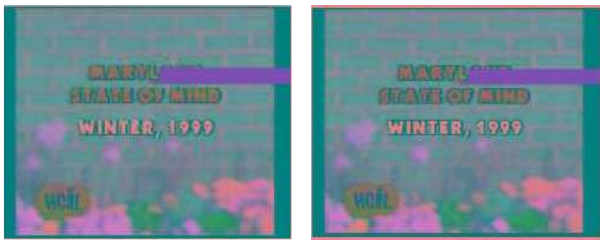


Fig. 9: Lab colour space frames

Here Xn, Yn and Zn are the tristimulus values of the reference white. Then for each converted frame color histogram is constructed with bin size of 10 for each one in LAb. Based on the local maximum of each fame of the histogram the cluster is formed using fuzzy c means clustering algorithm. In this the similar colors are grouped together. Each one is considered as an object based on the color. The local maximum of the histogram is considered as the number of cluster for each frame. So that the colors in the frame are grouped into separate groups. For each frame the process is repeated and then the groups of each frame is compared to extract the similar object based on the color. Normalization process is performed. Figure 8 and 9 shows the RGB to Lab colour space converted frames.

Motion feature extraction: Motion feature of the video can be used for retrieval. It is nothing but the movement between two consecutive frames. Motion estimation is to detect how these similar parts move. A motion vector is defined to represent the movement. It has a format (x, y). X is the distance (how many pixels) it moves in x direction, while y is the distance in y direction. Different algorithms are there to find out motion between two frames. In this work motion feature is extracted by comparing the block of two frames. A video frame will be divided into several blocks. Each block will be processed independently. For each block called current block, we try to match it with some similar parts of other frames called reference frames. The lower difference means more similar between two blocks. Motion estimation is to detect how

these similar parts move. It is performed using the following formula Eq. 3:

$$\sum_{x=1}^{x=m}\sum_{y=1}^{y=n}(C(x,y)-R(x,y))2 \quad (3)$$

Where C(x,y) and R(x,y) are current and reference blocks respectively.

SIFT feature: Sift is used to extract features from the frames (Lowe, 1999). Sift gives 128 dimensional feature vector. The features of the frames are clustered using Fuzzy c means clustering technique.

The four stages in the sift are:

- Scale-space extrema detection
- Keypoint localization
- Orientation assignment
- Generation of keypoint descriptors

Scale-space extrema detection: This stage of the Sift attempts to identify those locations and scales that are identifiable from different views of the object. The scale space of an image I(x,y) can be obtained by convoling the image with a variable scale Gaussian G(x, y, σ) Eq. 4:

$$L(x,y,\sigma)=G(x,y,\sigma)*I(x,y) \quad (4)$$

where, * is the convolution operator. Stable keypoints are detected by the difference of Gaussians technique. It is performed by a simple image subtraction of two nearby scales separated by a constant multiplicative factor k Eq. 5:

$$D(x,y,\sigma)=L(x,y,k\sigma)-L(x,y,\sigma) \quad (5)$$

Keypoint Localization: In this phase the low contrast keypoints and the keypints which are poorly aligned on edges are eliminated. At each candidate location, this is done by using the Taylor series expansion of scale space function Eq. 6:

$$D(x)=D+\frac{1}{2}(\frac{\partial D}{\partial x})^{T}x\frac{1}{2}x^{T}\frac{\partial^{z}D}{\partial x^{z}_{x}} \quad (6)$$

where, D and its derivatives are evaluated at the sample point and x= (x, y, σ)^T is the offset from the point. The location of the extremum is obtained by taking the derivative of this function with respect to x and setting it to zero Eq. 7:

$$X=(\frac{\partial^{z}D}{\partial x^{z}})-1\frac{\partial D}{\partial x} \quad (7)$$

The function value at the interpolated extremum is calculated by placing the offset value in Taylor series expansion. It is given as Eq. 8:

$$D(x)=D+\frac{1}{2}(\frac{\partial D}{\partial x})^{T}x \quad (8)$$

If the function value at this extremum is below a threshold value then this point is excluded. This eliminates extrema with low contrast.

Orientation Assignment: In this phase orientation is assigned to each keypoint based on local gradient properties. The scle of the keypoint is used to select the Gaussian smoothed image L. For each image sample, L(x, y) at this scale the gradient magnitude, m (x, y) and orientation, θ (x, y) is precomputed using pixel differences Eq. 9 and 10:

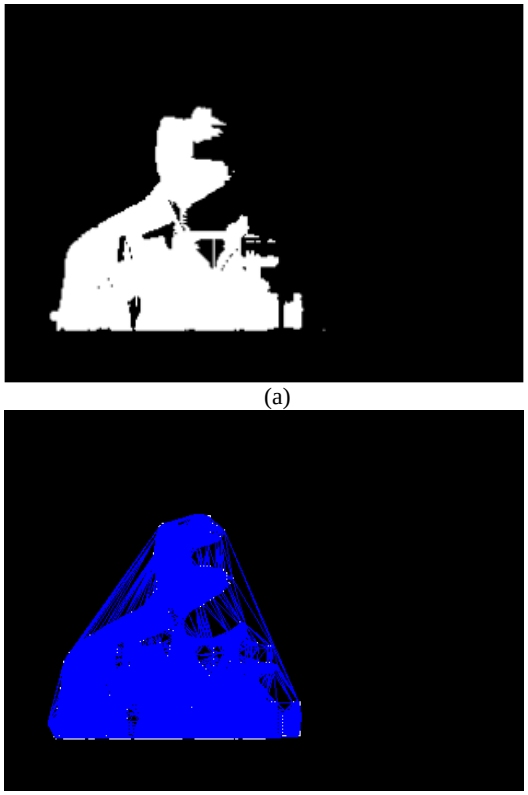
$$m(x,y)=\sqrt{(L(x+1,y)-L(x-1,y))^{2}+L(x,y+1)-L(x,y-1)^{2}} \quad (9)$$

$$\theta(x,y)=\tan^{-1}\left(\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right) \quad (10)$$

The orientation histogram is formed from the gradient orientation of sample points. Highest peak in the histogram is obtained to create keypoints with that orientation. Some of the key points will be assigned multiple orientations.

Keypoint Descriptor: In this stage the descriptor is computed which is invariant to affine distortions and change in illumination this descriptor is a 8bin orientation histogram representing 8 major directions. One such 8 bin vector is generated for each 4*4 sub region around the key point. Thus the total length of descriptor vector is 128.

Shape Feature: The shape of objects plays a powerful role among the different image features in content based similarity search and retrieval systems. Triangular mesh based shape description technique is implemented in this work. The basic idea in our work is to form the triangular mesh over the binary object. Figure 10a shows the Binary object from the frame. The total number of triangle meshes formed is considered as the shape feature. Figure 10b shows the mesh formation over the binary query object.



(a)

(b)

Fig. 10: (a) Binary object (b) Formed triangles

Retrieval: To retrieve the similar frames the object is given as the query. For the query object also all the above said features are extracted and are compared with features in the database. In this work kullback leibur distance measure is used for comparison.

RESULTS AND DISCUSSION

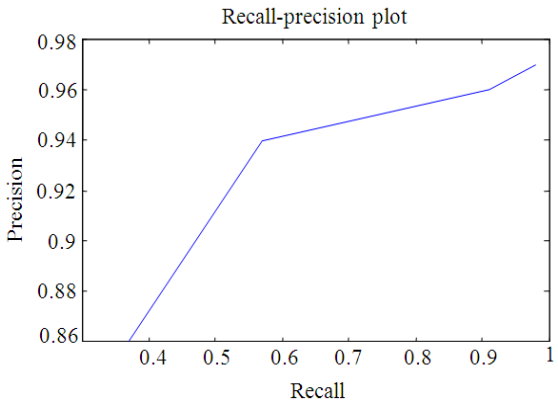
The proposed method is applied on 50 different videos. For different input query the relevant frames are retrieved from the database. We are getting the better results even after applying some rotation and scaling. In CBVR the performance measures are Precision and Recall. Graph-1 shows the Precision-Recall plot. Precision and recall are defined as:

$$\text{Precision} = \frac{\text{number of relavant images retrieved}}{\text{total images retrieved}}$$

$$\text{Recall} = \frac{\text{number of relavant images retrieved}}{\text{total number of relavant images}}$$

CONCLUSION

From the obtained results it is clear that the combination of the proposed features performs well in shot detection and CBVR. It is robust to rotation, scaling and illumination changes. The video is intiallly segmented using genetic edge histogram and DCT method. The kullback liebler distance similarity measure is used for comparison between the features in the feature library and the features of the query object extracted in a similar manner. The kullback Leibler distance measure gives efficient retrieval.



REFERENCES

Anjulan, A. and N. Canagarajah, 2006. Video scene retrieval based on local region features. Proceedings of the IEEE International Conference on Image Processing, Oct. 8-11, IEEE Xplore Press, Atlanta, GA., pp: 3177-3180. DOI: 10.1109/ICIP.2006.313044

Anjulan, A. and N. Canagarajah, 2007. Object based video retrieval with local region tracking. Image Commun., 22: 607-621. DOI: 10.1016/j.image.2007.05.008

Flickner, M., H. Sawhney, W. Niblack, J. Ashley and Q. Huang *et al*., 1995. Query by image and video content: The QBIC system. IEEE Comput., 28: 23-32. DOI: 10.1109/2.410146

Gunsel, B. and A.M. Tekalp, 1998. Content-based video abstraction. Proceedings of the International Conference on Image Processing, Oct. 4-7, IEEE Xplore Press, Chicago, IL, pp: 128-132. DOI: 10.1109/ICIP.1998.727150

Jain, A.K., 1989. Fundamentals of Digital Image Processing. 1st Edn., Prentice Hall, New Jersey, ISBN-10: 0133361659, pp: 569.

Kim, J.H., H.Y. Lim and D.S. Kang, 2008. An implementation of the video retrieval system by video segmentation. Proceedings of the 14th Asia-Pacific Conference on Communications, Oct. 14-16, IEEE Xplore Press, Tokyo, pp: 1-5.

Lowe, D.G., 1999. Object recognition from local scale-invariant features. Proceedings of the 7th IEEE International Conference on Computer Vision, Sept. 20-27, IEEE Xplore Press, Kerkyra, pp: 1150-1157. DOI: 10.1109/ICCV.1999.790410

Smith, J.R. and S.F. Chang, 1996. VisualSEEk: A fully automated content-based image query system. Proceedings of the 4th ACM International Conference on Multimedia, Nov. 18-22, ACM Press, Boston, MA, USA., pp: 87-98. DOI: 10.1145/244130.244151

Thakre, K.S., A.M. Rajurkar and R. Manthalkar, 2010. An effective CBVR system based on motion, quantized color and edge density features. Proceedings of the 1st International Conference on Intelligent Interactive Technologies and Multimedia, Dec. 28-30, ACM Press, Allahabad, India, pp: 145-149. DOI: 10.1145/1963564.1963589

Zhong, D. and S.F. Chang, 1997. Video object model and segmentation for content-based video indexing. Proceedings of the IEEE International Symposium on Circuits and Systems, Jun. 9-12, IEEE Xplore Press, pp: 1492-1495. DOI: 10.1109/ISCAS.1997.622202