CONTENT-BASED VIDEO PROCESSING USING SCENE SEGMENTATION

M.S. Sowmya[#1], J. Sriranjani[#2], V. Sujatha[#3], P V Rajaraman[*4]
[#] *Computer Science Engineering,*
Rajalakshmi Engineering College, Thandalam, Chennai, Tamil Nadu, India.
[*] *Computer Science Engineering,*
Rajalakshmi Engineering College, Thandalam, Chennai, Tamil Nadu, India.
[1]sowmya0430@gmail.com
[2]sriranjanijagadeesan@gmail.com
[3]sujathaveeswar@gmail.com
[4]rajaraman.pv@rajalakshmi.edu.in

***Abstract*- The amount of video data is continuously growing. A video has a huge amount of raw and dynamic data which is of much richer content than individual images. This inspires the researchers world-wide. This paper will provide information about the three approaches in scene segmentation-Key frame-based approach, audio and Vision Integration-Based approach, Background-Based Approach, their algorithms, architectures and accuracy measures. Finally we investigate the future enhancements.**

***Index Terms*- Video Structure analysis, Scene Segmentation, Key frame-based approach, audio and Vision Integration-Based approach, Background-Based Approach.**

I. INTRODUCTION

A video is a Visual multimedia source that is a sequence of images forming a moving picture. A video has many applications such as surveillance, education, and many others which has increased the researcher's interest in finding the various functions that can be formed using the videos. A video has two channels- auditory and visual. The information available in a video are of four types [16]. One, Audio information in the auditory channel. Two, Visual information such as images contained in the visual channel. Three, Transcripts obtained from speech transcripts and caption text and video metadata which are tagged texts including title, texts, etc.

A video is divided into parts as specified in the hierarchy [1]. It is subdivided into scenes and then shots which are further subdivided into frames as shown in Fig1.

A frame is a still image. A shot is a single series of action using one camera. A scene is a group of shots and is similar to a paragraph in a text. A video is a group of scenes.

The aim of video structural analysis is to segment the video into its structural parts which have semantic contents [2], segmentation of scene and boundary detection of shot and extraction of the frame [4]. In this paper, we are going to concentrate on scene segmentation.

Scene segmentation is also called story unit segmentation. A group of continuous shots that are coherent in theme and subject is called as a scene. Scenes have higher semantics than a shot. Scenes are segmented by identifying shots with similar content. The grouping may use the information present in texts, images or the audio track in the video [15].

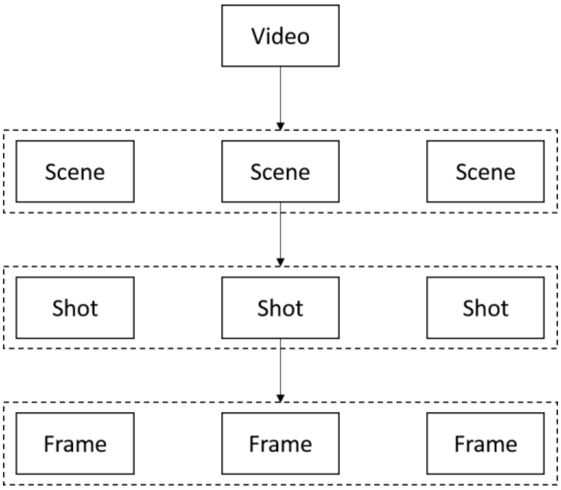

Fig. 1 General Hierarchy of Video Parsing

The scene segmentation approaches can be divided into four categories according to the processing method [18]. These are 1) Merging Based, 2) Splitting Based, 3) Statistical model-Based, 4) Shot boundary Classification.

A. *Merging- Based approach*

It performs a bottom-up process to group different shots to scenes. It is a two pass scene segmentation process. Backward shot coherence is used to carry out oversegmentation of scenes in the first pass. Merging of these oversegmented scenes which are identified using motion analysis is done in the second pass. There is a previously proposed best model for merging [20]. The algorithm takes each shot as a hidden state and loops upon the boundaries between consecutive shots by a left-right HMM [2].

B. *Splitting-Based approach*

It uses a top-down style thus splitting the whole video into coherent scenes [17]. A similarity graph was constructed for a video in which normalized cuts were used to divide the graph into sub graphs. These sub graphs represent the scenes in a video. Previous researches have introduced a scene definition for narrative films [21]. They also presented a technique to cluster them into relevant shots forming a scene which uses that definition.

C. *Statistical model-Based approach*

In this approach, a statistical model of shots is created to segment the scenes. Each scene is modelled with a Gaussian density, A research already defines a unified energy minimization framework in which global content constraint between individual shots and local temporal constraint between adjacent shots are both represented. A boundary voting procedure decides the optimal scene boundaries [22].

D. *Shot boundary classification*

The scene and the non –scene boundaries are classified, shot boundary's features are extracted. A research presents a genre-independent method to detect scene boundaries in broadcast videos [23]. It is based on classification of scene and non-scene changes. SVM is used to classify the shot boundaries. Hand-labelled video scene boundaries from a variety of broadcast genres, which are used to generate positive and negative training samples for SVM.

There are three groups of approaches to perform scene segmentation. These groups are performed according to the representation of shots. The three groups are 1) key frame-Based approach, 2) Audio Vision Integration-Based approach and 3) Background- Based Approach which is going to be explained in Sections following. These are the processing types that can be used.

II. Key Frame-Based Approach

Frames of the same shot contain a great amount of redundancy. Certain frames that best reflect the contents of the shot are selected as key frames. They must be selected in such a manner to reduce the redundancy [5]. The features used to extract key frames include colours (particularly histograms [3].), edges, optical flow, etc.

Temporally close shots are considered to belong to the same scene [7]. Sometimes, the similarities among shots are calculated using contrasting the key frames [24]. Same shots are linked by overlapping links. These scenes are segmented. Motion trajectories are also extracted and analysed and then encoded to form images of large volumes of temporal slices. A motion based-Key frame is used to efficiently represent the shot contents. The similarity between different shots are measured to detect the scene changes.

The limitation of key frame-based approach is that, dimensions of the shot contents are not efficiently represented by key frames. This is because, shots within a scene are correlated by dynamic contents within the scene rather than by key frame-based similarities between shots.

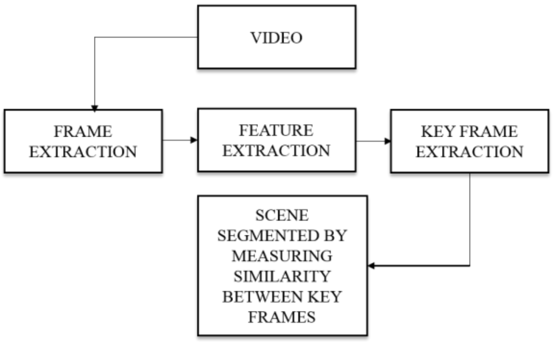

Fig. 2 Key Frame Extraction

III. Audio Vision Integration Based Approach

This approach selects a scene boundary based on the simultaneous visual and audio content changes [10]. It is possible to detect audio scenes and video scenes separately. A time-constrained nearest neighbour algorithm [8] is used to determine the correspondences between the frames. The algorithm takes input as test and training data.

The test data contains all the frames belonging to a particular video and the train data contains the frames, usually the key frames [9] to find the correspondence between them to each of the test data. Frames with highest correspondence are grouped into a scene. The limitation of this approach is that it is difficult to determine the relation between audio segments and visual shots [1].

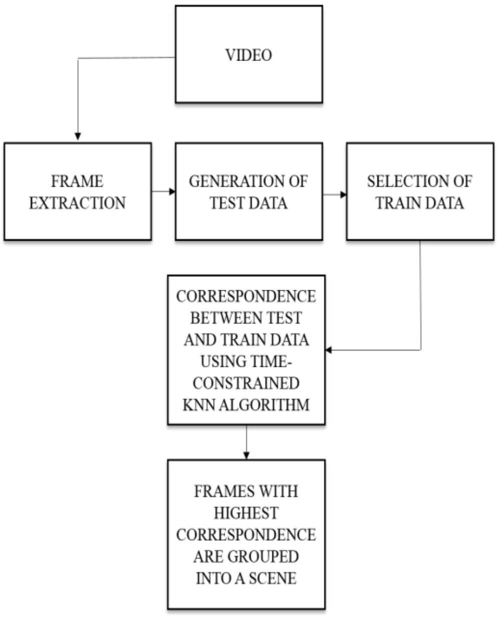
VIDEO
FRAME EXTRACTION
GENERATION OF TEST DATA
SELECTION OF TRAIN DATA
CORRESPONDENCE BETWEEN TEST AND TRAIN DATA USING TIME-CONSTRAINED KNN ALGORITHM
FRAMES WITH HIGHEST CORRESPONDENCE ARE GROUPED INTO A SCENE

Fig. 3 Audio Visual Integration Approach

IV. Background Based Approach

This approach is based on the concept that a particular frame belongs to a scene based on the similarity that they have in a background [12]. The classification of frames into scenes is done with the change in background [14]. A frame in scene contains certain background which remains constant for the next frame also [1].

This will help in detecting the next scene of the video sequence. The background can be subtracted from all scene and the changes in the frames can be determined [11] on a measure of how different the frame is from the background. The difference in each frame determines the class of scenes and shots in the video that is being analysed [13].

The boundary detection is done using shot boundary detection method where the frames are split into different scenes. There is a classification of the frame into different shots using their differences instead of similarities [12]. Shots are represented using the difference in each key frame where as other approaches such as merging, splitting and statistical model uses the similarities in the key frames. The usage of difference in key frame makes shot boundary methods better than the other.

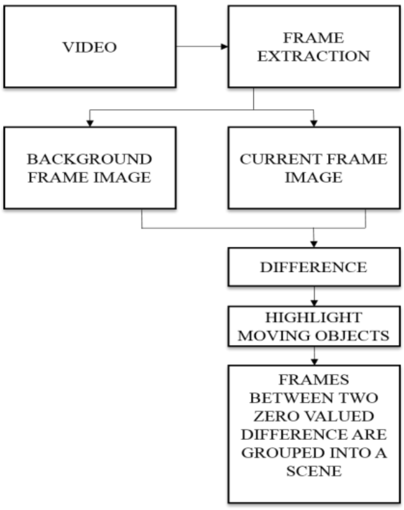

Fig. 4 Background Based Approach

V. Algorithm

The algorithm for the different approaches are explained in this section.

The algorithm for key frame-based approach proceeds as follows. The video to be segmented into scenes is given as the input. The processing performs the following steps.

1) Two frames are considered at each time and their difference in histograms [3] are calculated.
2) The mean and variance are calculated as per the formula

$$\text{Mean} = \frac{(\sum_{i=1}^{n} x_i)}{n}$$

$$\text{Standard Deviation (SD)} = \sqrt{\frac{(\sum_{i=1}^{n}(x_i - \bar{x})^2)}{n-1}}$$

(Or)

$$\sqrt{\frac{(\sum_{i=1}^{n}(x_i - \bar{x})^2)}{n}}$$

3) The threshold value is calculated.
4) Then, any consecutive pair of frames whose absolute difference is greater than the threshold value is selected as a keyframe [6] and as a scene change.

The output of the procedure are a set of keyframes which defines the scene change.

The Audio-Vision integration uses the following algorithm. The input to this is the video and a set of training images. The training images are selected according to the factor based on which the video should be segmented. For, example characters can be selected. The following procedure is followed.

1) The frames from the video are extracted and stored as test images.
2) Each training image has a class/group that indicates it.
3) Each test image is compared with all the training set and the best match is found using algorithms such as a time-constrained nearest neighbour algorithm [8], and that specific class is assigned to that frame.
4) Scenes are segmented based on the assumption that whenever the class of the frame changes, there is a change in the scene.

The output of this algorithm is the test data that consists of every frame in that video and also the classes that each frame is segregated into.

The algorithm for background based-approach is as follows. The input for this approach is the video and an image containing the background of that video's shots.

1) Compare each frame with the background. Mark any changes present using any indication.
2) Check the indication and the scenes containing the background are identified
3) Scenes are segmented using the assumption that, from one point where background appears to the next consists of one scene.

The output of this algorithm consists of the indications for every frame whether or not background appears. And, finally the scenes are segmented using these indication.

VI. Future Developments

Although there are many developments in the processing of video, content based video indexing still has many approaches to be explored. Processing of any video starts with structural analysis. For further processing of video through content based retrieval and analysis the following developments can be carried out.

A. Video Data mining

Data mining uses the key features of structural analysis to get data details of video such as subtitle file, audio, time frame, scenes, shots etc.

B. Video Classification

Categorized Classification of video is done to improve the efficiency of video retrieval. This will help in the performance of the retrieval and indexing.

C. Video Annotation

Video annotation is also done to improve performance and efficiency. It is similar to classification but it has a concept of allocating videos into different semantics of a video.

D. Indexing using Query and retrieval

Different types of queries are used to retrieve videos which is a process of indexing. Querying can be done using objects, frames, text search, examples, sketches etc. Retrieval is done by different methods of matching.

All the above developments are used for indexing videos by browsing so that there are no duplicates formed while indexing the video.

VII. CONCLUSION

We have presented the implementation of structural analysis of video which is the first step towards content based video indexing. Structural analysis of video was done using shot boundary detection, key frame extraction and scene segmentation. Scene segmentation was executed using three approaches: key frame based approach, audio-vision integration based approach and background based approach. This analysis is used as the base for implementation of video indexing and retrieval such as video data mining, video classification and annotation.

ACKNOWLEDGEMENT

We would like to express our deep sense of gratitude to our project guide. Mr.Rajaraman PV, for his constant encouragement and kind support and for providing the necessary guidance throughout this project work.

REFERENCES

[1] W. Hu, N. Xie, L. Li, X. Zeng and S. Mayback, “A survey on visual content-based video indexing and retrieval,” IEEE Trans. Systems, Man and Cybernetics, vol. 41, no. 6, pp. 798-803, Nov. 2011.

[2] M.N. Asghar, F. Hussain and R. Manton, “Video indexing: a survey,” IJCIT, vol. 03-issue 01, Jan 2014.

[3] A.V. Kumthekar and J.K. Patil, “Key frame extraction using color histogram method,” IJSRET, vol. 02-issue 04 pp. 207-214, Jul 2013.

[4] G. Liu and J. Zhao, “Key frame extraction from mpeg video stream,” proceedings of the second symposium international comp. science and computational tech. pp. 007-011, Dec 2009.

[5] S.D. Thepade and A.A. Tonge, “An optimized key frame extraction for detection of near duplicates in content based video retrieval,” IEEE Trans. Comm. and signal Proc. Conf., pp. 1087-1091, Apr 2014.

[6] R. Pan, Y. Tian and Z. Wang, “Key frame extraction algorithm based on entropy,” IEEE Int. Conf. ICEEE, pp. 001-004, Nov 2010.

[7] J. Rong, W. Jin and L. wu, “Key frame extraction using inter-shot information,” IEEE Int. Conf., Multimedia, vol. 01, pp. 571-574, Jun 2004.

[8] Y. Li and B. Cheng, “An improved k-nearest neighbour algorithm and its application to high resolution remote sensing image classification,” IEEE, pp. 001-004, Aug 2009.

[9] M. Zhang and Z. Zhou, “A k-nearest neighbour based algorithm for multi- label classification,” IEEE Int. Conf, vol. 02, pp. 718-721, Jul 2005.

[10] M. Yao, “Research on learning evidence improvement for knn based classification algorithm,” inter. journal, theory and application, vol. 07, pp. 103-110, 2014.

[11] Y. Benezeth, P.-M. Jodoin, B. Emile, H. Laurent and C. Rosenberger, “Comparative study of background subtraction algorithms,” journal of electronic imaging, Oct 2012.

[12] D. Das and S. Saharia, “Implementation and performance evaluation of background subtraction algorithms,”

[13] D.H. Parks and S.S. Fels, “Evaluation of background subtracting algorithms with post-processing,” IEEE Int. Conf. pp. 192-199, Sept 2008.

[14] X. Zou, Z. Chi and X. Zhao, “A robust background subtraction approach with a moving camera,” IEEE Int. Conf., pp. 1026-1029, Dec 2012.

[15] L.-H. Chen, Y.-C. Lai, and H.-Y.M. Liao, “Movie scene segmentation using background information,” pattern recognit., vol. 41, no. 3, pp. 1056-1065, Mar. 2008.

[16] H. Sundaram and S.-F. Chang, “Video scene segmentation using video and audio features,” in Proc. IEEE Int. Conf. Multimedia Expo., New York, 2000, pp. 1145-1148.

[17] Z. Rasheed and M. Shah, “Detection and representation of scenes in videos,” IEEE Trans. on Multimedia, vol. 7, no. 6, pp. 1097–1105, Dec. 2005.

[18] H.-W. Kang and X.-S. Hua, “To learn representativeness of video frames,” in Proceedings of the 13th annual ACM international conference on Multimedia. ACM, 2005, pp. 423–426.

[19] S. Russel and P. Norvig, “Artificial Intelligence – a modern approach,” second edition, Pearson Education, 2003.

[20] L. Zhao, W. Qi, Y.-J. Wang, S.-Q. Yang and H. Zhang, “Video shot grouping using best-first model merging,” in Proceedings of SPIE, vol. 4315, 2001, p. 262

[21] W. Tavanapong and J. Zhou, “Shot Clustering Techniques for Story Browsing.”, IEEE Transactions on Multimedia, vol. 6, no. 4, pp.517-527, Aug. 2004.

[22] Z. Gu, I.Mei, X.-S. Hua, X. Wu, and S. Li,”EMS: Energy Minimization Based Video Scene Segmentation”, Multimedia and Expo, 2007 IEEE International Conference on, pp.520-523,Jul. 2007.

[23] N. Goela, K. Wilson, F. Niu, A. Divakaran and I. Otsuka, “An SVM Framework for genre-independent scene change detection”, in Multimedia and Expo, 2007 IEEE International Conference on. IEEE, 2007,pp. 532-535

[24] A. Hanjalic, R. L. Lagendijk and J. Biemond, “Automated high-level movie segmentation for advanced video-retrieval systems,” Circuits and Systems for Video Technology, IEEE Transactions on, vol. 9, no. 4, pp. 580-588, 1999.