International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Content-Based Video Indexing and Retrieval using Key frames Texture, Edge and Motion Features

M.Ravinder†* and T.Venugopal‡

†JNTUK, Kakinada, Andhra Pradesh, India
‡Department of CSE, JNTUHCES, Sultanpur, Medak, Telangana, India

Accepted 25 April 2016, Available online 30 April 2016, **Vol.6, No.2 (April 2016)**

Abstract

In this paper, a novel algorithm for content-based video indexing and retrieval using key-frames texture, edge, and motion features is presented. The algorithm extracts key frames from a video using k-means clustering based method, followed by extraction of texture, edge, and motion features to represent a video with the feature vector. The algorithm is evaluated on a database of three hundred and thirty five videos (collected from TRECVID 2005, Google, and BBC) of four types. The performance of the proposed framework is compared with volume local binary patterns (VLBP) method. The proposed algorithm outperforms well compare to VLBP method.

Keywords: *Video, indexing, retrieval, texture, edge, motion, frame.*

1. Introduction

The availability of internet, quality recording, and huge storage multimedia technologies with low cost, allows a user to record a video and store it in a video repository. There is an increase in demand, for an efficient annotation and retrieval framework to maintain the huge video repositories. A number of content-based video indexing and retrieval frameworks have been proposed in the literature. The initial step in content-based video indexing and retrieval (CBVIR) framework is, dividing a video in to smaller portions known as video shots. A video shot consists of a sequence of similar content frames (Weiming Hu *et al*., 2011). The shot boundary detection algorithms are helpful in separating a video in to video shots. The main steps engaged in a shot boundary detection framework are, the first step is extracting features from each individual frame of the video, then, compare the feature vectors of the frames, to find the shot boundaries. Distinctive features used for shot boundary detection are, color histogram features (C. H. Hoi, L. S. Wong, and A. Lyu, 2006), motion features (S.V.Porter, 2004), edge change ratio features (Z.-C. Zhao and A.-N. Cai, 2006), and corner points (X. B. Gao, J. Li, and Y. Shi, 2006). The next step in CBVIR framework is representing a video shot with the features took from it. One of the methods employed to represent a video shot is, extraction of key frame followed by representing the video shot with the features of the key frame. Unique key frame extraction methods proposed in the literature can be noted in (R.Hamid *et al*., 2007; G. Lavee *et al*., 2009; J. Tang *et al*., 2009; X. Chen *et al*., 2009).

The authors of (Bart Thomee *et al*., 2015) introduced a benchmark dataset and discussed the future challenges of multimedia data. One of the statements given by the authors is that, there exist huge amount of image and video data with less metadata and erroneous annotations, and they claimed that, proper annotation of the multimedia data as one of the future challenges. In (Fillipe Souza *et al*., 2015), the authors have introduced a new pattern based video understanding method. A natural language description of video segments has been introduced in (Niveda Krishnamoorthy *et al*., 2013). The authors of (Pradipto Das *et al*., 2013) have proposed a method useful for describing a video segment with thousand frames with few words. A content and concept based video retrieval framework is introduced in (Jeffrey Dalton *et al*., 2013). In (D.Sudha *et al*., 2015), the authors have presented a survey on different algorithms that are useful for video retrieval by reducing the semantic gap between high level and low level features. The authors of (Muhammad Nabeel Asghar *et al*., 2014) have presented a survey on different video indexing frameworks. In (Hatim G. Zaini *et al*., 2014) have introduced a content-based video retrieval framework using multiple features and semantic concepts. In (Matthijs Douze *et al*., 2010), a novel method based on individual frame comparisons between query video and videos in the video database has been introduced. A framework based on one dimensional video distance trajectories has been proposed in (Zi Huang *et al*., 2010). The authors of the paper (Shangfei Wang *et al*.,

*Corresponding author **M.Ravinder** is a Research Scholar of CSE **and Dr.T.Venugopal** is working as Associate Professor

2015) have presented a survey on existing video content analysis frameworks and projected the future challenges.

Textures with motion are known as temporal textures (Szummer, M., and Picard, R.W, 1996). In (Szummer, M. *et al*.,1996), the authors have introduced a spatial temporal auto aggressive model useful for temporal texture recognition. In (Lifeng Shang *et al*., 2010), the authors have proposed two new frameworks for near duplicate video retrieval based on spatiotemporal features using conditional entropy and local binary patterns. (Chetverikov.D, and Peteri.R, 2005), have put forward a survey on dynamic textures, and five types of extraction methods. (Fazekas.S *et al*, 2005), have compared the dynamic texture classification methods based on normal flow features, and complete flow features. (Smith.J.R *et al*., 2002), have proposed a novel method handy for video indexing based on spatio-temporal wavelets. In (Ja-Hwung Su *et al*., 2009), the authors have proposed a method for content-based video retrieval based on temporal patterns mining mechanism. A survey on existing frameworks of near duplicate video retrieval and future challenges have been presented in (Jiajun Liu *et al*. 2013).

2. Related Work

The texture based methods discussed in this section are, local binary patterns (LBP), and volume local binary patterns (VLBP).

2.1 Local binary patterns (LBP)

(Ojala *et al*., 1996) have introduced the popular and successful method of texture extraction, called local binary patterns (LBP). The LBP value for a pixel centered in a 3x3 patch of an image can be found by comparing the value of it with the neighborhood pixels, if the neighborhood pixel value is greater than center pixel value, replace the neighborhood pixel value with one, and otherwise replace with zero. By multiplying the resultant 3x3 patch values with corresponding binary weights and sum of products yields the LBP value for the center pixel as shown in figure, Fig.1.

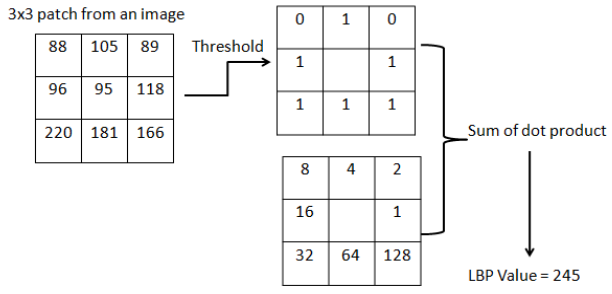

Fig. 1. Example calculation of LBP value for a 3x3 patch

The same procedure is implemented in all the 3x3 patches of the image, which yields the texture image. The histogram of texture image serves as the LBP feature vector of the image.

2.2 Volume local binary patterns (VLBP)

(Guoying zhao *et al*., 2006), have extended the local binary patterns method, and represented a video with dynamic textures. Volume local binary patterns method mainly based on extracting dynamic textures from a sequence of three frames at a time. An example computation of volume local binary patterns for a 3x3x3 cubic patch is as showed in figure, Fig.2.

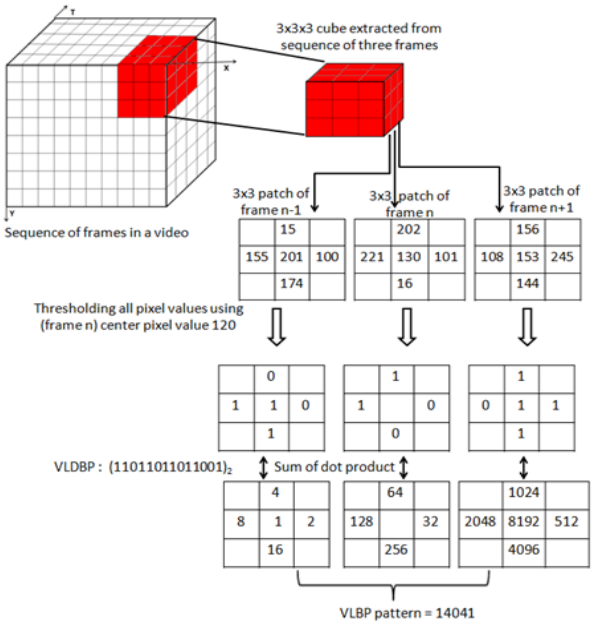

Fig.2. An example calculation of VLBP pattern value for a 3x3x3 cubic patch

Applying threshold operation on four neighborhood pixels of 3x3x3 cubic patch, then, sum of dot product with the corresponding binary weights will yield the volume local binary pattern value for the center pixel.

3. Proposed algorithm

In this section, the novel algorithm meant for content-based video indexing and retrieval is introduced.

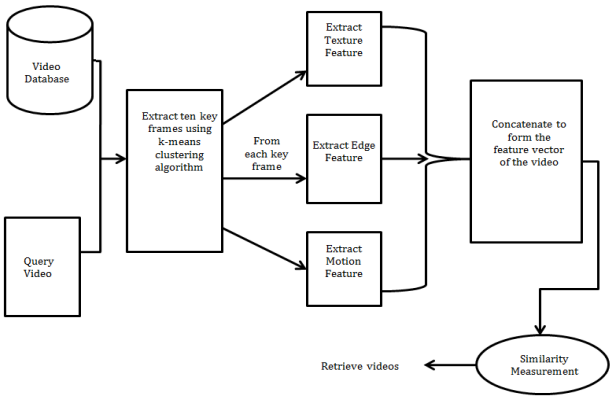

Fig.3. Proposed framework

The framework of the proposed algorithm is as depicted in figure, Fig.3.

The first step in our framework is, from the input video selecting ten key frames using k-means clustering algorithm. The second step is extracting texture, edge, and motion features from each of the key frames, after extracting the features from all the ten key frames the next step is concatenation of all the features extracted to form the feature vector of the video and index the video with the feature vector.

3.1 Key-frames extraction using k-means clustering algorithm

The first step in our proposed framework is extraction of ten key frames using k-means clustering algorithm. The k-means clustering algorithm is useful to quantize the n data items in to k groups of similar data items. A set of data item (d1, d2, d3, ……., dn) are given, where all the data items are vectors of same size.

The k-means clustering algorithm consists of following steps:

Step 1: Initialize k cluster center centers with the randomly selected k data items from the given set of data items.

Step 2: Assign the data items to their corresponding nearest cluster centers.

Step 3: Find the mean of each cluster of data items, and update the new cluster centers with the mean of corresponding cluster of data items.

Step 4: Continue step 2 and step 3 until convergence of cluster centers.

By using the k-means clustering algorithm as explained above, the steps involved in finding the key frames from the given input video are:

Step 1: From the input video extract the color histogram of each frame and store them as set of feature vectors.

Step 2: Apply k-means clustering algorithm and partition the set of feature vectors to ten clusters.

Step 3: From a cluster of feature vectors, find out the feature vector which is near to the corresponding cluster center.

Step 4: After that declare the corresponding frame of the feature vector as key frame.

Step 5: From all the ten clusters extract ten key frames by applying step 3 and step 4 on each cluster.

3.2 Texture, edge, and motion features extraction

The next step in our proposed framework is extraction of texture, edge, and motion features from the key frames extracted in the previous step.

Texture, edge, and motion features of each key frame are extracted by considering one frame at a time. First we will apply the texture, edge, motion feature extraction methods on first key frame after that the same procedure is followed on all the remaining nine key frames.

The procedure followed to extract the texture features is:

Step1: Divide the frame in to four equal parts (approximately).

Step 2: On each part of the frame extract local binary patterns.

Step 3: Concatenate all the features extracted in step 2, to form the texture feature vector of the key frame.

The procedure followed to extract the edge features of a key frame is:

Step 1: On the input key frame apply the two sobel filters (3x3 filters) which results in two images (X, Y).

Step 2: By using the resultant images of step 1, find the direction of each pixel by using $\tan^{-1}(Y/X)$.

Step 3: Find the histogram (edge features) of resultant image of step 2.

The procedure followed to extract motion features of a key frame is:

Step 1: For the input key frame kf find either kf-10 th frame (if kf+10 > Number of frames of the video), or kf+10 th frame.

Step 2: Find the absolute difference between two frames.

Step 3: Find the histogram (motion features) of resultant frame of step 2.

After extracting texture, edge, and motion features of all the ten key frames, the feature vector of the video is formulated by concatenating all the extracted features of ten key frames.

The same procedure is applied to all the videos in the database, and for a query video. The resultant feature vector of the query video is compared with the feature vectors of the videos in the video database, with the help of Euclidean distance measurement algorithm, which results in retrieval of videos from the database.

4. Experimental results

To implement the proposed algorithm, a video data set of three hundred and thirty five videos, in which seventy two boat, eighty car, and one hundred and forty eight air-plane, and thirty five war tank videos are there. The data set is available at (http://vision.eecs.ucf.edu/projects/arslan/vidmatching/index.htm).

The experimental results of our algorithm compared with existing method volume local binary patterns (VLBP). Precision and recall measures are used to evaluate the retrieval performance, which are calculated using the following equations (2), and (3).

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (2)$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (3)$$

The experimental results of the proposed framework, and the existing method VLBP, are as shown in table 1, and table 2.

Table 1 Precision (N=5) (%) (Top 5 retrieved videos)

Category	VLBP	Proposed Framework
Boats	46.94	52.5
Cars	47.75	60.5
Air planes	58.78	75.68
War tanks	36	44.57
Average value	47.3675	58.3125

Table 2 Recall (N=35) (%) (Top 35 retrieved videos)

Category	VLBP	Proposed Framework
Boats	16.98	15.24
Cars	15.36	19.08
Air planes	10.57	14.38
War tanks	15.97	19.84
Average value	14.72	17.135

The average precision, and average retrieval rate, of our proposed method and the existing VLBP method, are graphically as shown in below figures, Fig.4, and Fig.5.

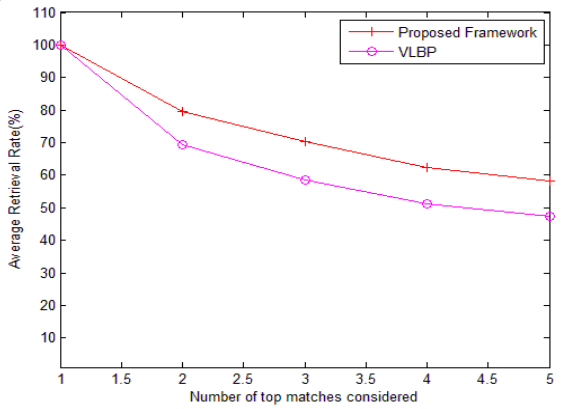

Fig.4. Comparison of proposed method with existing method VLBP in terms of average precision

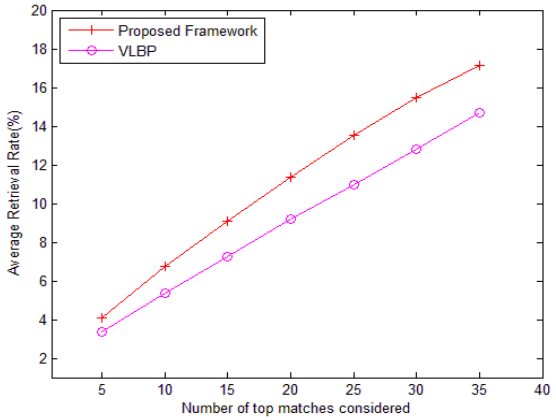

Fig.5. Comparison of proposed method with existing method VLBP in terms of average retrieval rate

Conclusions

In this paper, a novel framework has been introduced based on key frames texture, edge, and motion features, meant for content-based video indexing and retrieval. The algorithm is tested on the data set of three hundred and thirty five videos and the performance of the proposed framework compare to existing method VLBP is reasonably good.

Acknowledgement

The authors want to acknowledge Arslan Basharat, for providing the dataset and made it available at (http://vision.eecs.ucf.edu/projects/arslan/vidmatching/index.htm).

References

Weiming Hu., Nianhua Xie., Li Li., Xianglin Zeng., and Stephen Maybank (2011). A Survey on Visual Content-Based Video Indexing and Retrieval. IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 41, No. 6.

G. Lavee, E. Rivlin, and M. Rudzsky. (Sep., 2009), Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 39, no. 5, pp. 489–504.

J. Tang, X. S. Hua, M. Wang, Z. Gu, G. J. Qi, and X. Wu. (Apr., 2009), Correlative linear neighborhood propagation for video annotation. IEEE Trans. Syst., Man, Cybern., B, Cybern., vol. 39, no. 2, pp. 409–416.

X. Chen, C. Zhang, S. C. Chen, and S. Rubin. (Mar. 2009), A human-centered multiple instance learning framework for semantic video retrieval. IEEE Trans. Syst, Man, Cybern., C: Appl. Rev., vol. 39, no. 2, pp. 228–233.

Ojala. T, Pietikäinen.M, Harwood.D (1996), A comparative study of texture measures with classification based on feature distributions. Pattern Recognition 29, 51-59.

Haoran.Yi, Deepu Rajan, Liang-Tien Chia (2005), A new motion histogram to index motion content in video segments. Pattern Recognition Letters 26, 1221–1231.

Muhammad Nabeel Asghar, Fiaz Hussain, and Rob Manton (2014), Video Indexing: A Survey, International Journal of Computer and Information Technology, Volume 03, Issue 01.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li (2015), The New Data and New Challenges in Multimedia Research. In arXiv:1503.01817v1 [cs.MM].

Hatim G. Zaini, and T. Frag (2014), Multi feature content based Video Retrieval Using High Level Semantic Concepts, IPASJ International Journal of Computer Science (IIJCS), Volume 2, Issue 9.

Matthijs Douze, Herve Jegou, Cordelia Schmid (2010), An image-based approach to video copy detection with spatio-temporal post-filtering, IEEE Trans. Multimedia 12, 4, 257–266.

Zi Huang, Heng Tao Shen, Jie Shao, Bin Cui, Senior Member, and Xiaofang Zhou (2010), Practical Online Near-Duplicate Subsequence Detection for Continuous Video Streams, IEEE Transactions on Multimedia, Vol.12, No.5.

Jiajun Liu, Zi Huang, Hongyun Cai, Heng Tao Shen,Chong Wah Ngo, and Wei Wang (2013), Near-Duplicate Video Retrieval: Current Research and Future Trends, ACM Computing Surveys, Vol. 45, No. 4, Article 44.

Shangfei Wang, and Qiang Ji (2015), Video affective content analysis: a survey of state-of-the-art methods, IEEE Transactions on Affective Computing, volume 6, Issue 4.

C. H. Hoi, L. S. Wong, and A. Lyu. (2006), Chinese university of Hong Kong at TRECVID 2006: Shot boundary detection and video search. in Proc. TREC Video Retrieval Eval.

S.V.Porter. (2004), Video segmentation and indexing usingmotion estimation. Ph.D. dissertation, Dept. Comput. Sci.,Univ. Bristol, Bristol, U.K.

Z.-C. Zhao and A.-N. Cai. (2006), Shot boundary detection algorithm in compressed domain based on adaboost and fuzzy theory. in Proc. Int. Conf. Nat. Comput., pp. 617–626.

X. B. Gao, J. Li, and Y. Shi. (2006), A video shot boundary detection algorithm based on feature tracking. in Proc. Int. Conf. Rough Sets Knowl. Technol., (Lect. Notes Comput. Sci.), 4062, pp. 651–658.

R.Hamid, S.Maddi,A. Bobick, and M. Essa. (Oct., 2007), Structure from statistics—Unsupervised activity analysis using suffix trees. in Proc. IEEE Int. Conf Comput. Vis., pp. 1–8.

Szummer, M., and Picard, R.W. (1996), Temporal texture modeling. In Proc. IEEE International Conference on Image Processing. Volume 3, 823-826

Chetverikov.D, Peteri.R (2005), A brief survey of dynamic texture description and recognition. In Proc. of 4th Int. Conf. on Computer Recognition Systems. Poland, 17-26.

Peteri. R., Chetverikov.D (2005), Dynamic texture recognition using normal flow and texture regularity. In Proc. Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2005), Estoril, Portugal, 223-230.

Fazekas.S, Chetverikov.D (2005), Normal versus complete flow in dynamic texture recognition: a comparative study. Texture 2005: 4th International Workshop on Texture Analysis and Synthesis, Beijing.

Smith.J.R, Lin.C.Y, Naphade.M (2002), Video texture indexing using spatiotemporal wavelets. In IEEE Int. Conf. on Image Processing (ICIP 2002). Volume 2, 437-440.

Zhao, Guoying, and Matti Pietikäinen (2006), Dynamic texture recognition using volume local binary patterns. WDV 2005/2006, LNCS 4358, pp.165-177.

Fillipe Souza, Sudeep Sarkar, Anuj Srivastava, and Jingyong Su (2015), Temporally Coherent Interpretations for Long Videos Using Pattern Theory. in CVPR2015.

Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond Mooney, Kate Saenko, and Sergio Guadarrama (2013), Generating natural-language video descriptions using text-mined knowledge, Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p. 10.

Pradipto Das, Chenliang Xu, Richard F. Doell, and Jason J. Corso (2013),A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2634–2641.

Jeffrey Dalton, James Allan, and Pranav Mirajkar (2013), Zero-shot video retrieval using content and concepts, in Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, pp. 1857–1860.

D.Sudha , and J.Priyadarshini (2015), Reducing Semantic Gap in Video Retrieval with Fusion: A survey, in 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15).

Ja-Hwung Su, Yu-Ting Huang, and Vincent S. Tseng (2009), Efficient Content-based Video Retrieval by Mining Temporal Patterns, ACM, MDM/KDD'08.

Lifeng Shang, Linjun Yang, Fei Wang,Kwok-Ping Chan, and Xian-Sheng Hua (2010), Real-time Large Scale Near-duplicate Web Video Retrieval, In Proceedings of the ACM Conference on Multimedia (MM'10). 531–540.