Indexing Video Database for a CBVCD System

Debabrata Dutta[1], Sanjoy Kumar Saha[2], and Bhabatosh Chanda[3]

[1] Tirthapati Institution, Kolkata, West Bengal, India,
[2] Computer Science and Engineering Department, Jadavpur University, Kolkata, India
[3] ECS Unit, Indian Statistical Institute, Kolkata, India
debabratadutta2u@gmail.com, sks_ju@yahoo.co.in,
chanda@isical.ac.in

Abstract. In this work, we have presented a video database indexing methodology that works well for a content based video copy detection (CBVCD) system. Video data is first segmented into cohesive units called shots. A clustering based method is proposed to extract one or more Representative frames from the shots. On such collection of representatives extracted from all the shots in the video database, triangle inequality based image database indexing scheme is applied. Thus, video indexing is mapped to the task of image indexing. For a shot, following the proposed methodology primarily candidate shots corresponding to the matched representative frames are retrieved. Only on such small number of candidates the rigorous video sequence matching technique can be applied to make final decision by the CBVCD system or video retrieval system. Experimental result with a CBVCD system indicates significant gain in terms of speed, reduces false alarm rate without much compromise in terms of correct recognition rate in comparison to exhaustive search.

Keywords: Video Database, Indexing, CBVCD.

1 Introduction

With the development of different multimedia tools and devices there has been enormous growth in the volume of digital video data. Search and retrieval of desired piece of video data from a large database has become an important area of research. In the applications like content based video retrieval (CBVR) or content based video copy detection (CBVCD) system a query video data is given by the user. It is to be matched with the video data stored in the database. In case of a CBVR system, we are mostly interested in finding few top order matches. But in case of CBVCD detection, the task is to decide whether or not the query is a copied version of any reference video data stored in the database. Further challenge for a CBVCD system is that a copied video data may undergo different photometric and post-production transformations making it different from the corresponding reference video. In both the applications, exhaustive video sequence matching is prohibitive. As a result indexing scheme becomes essential. In this work, we propose an indexing scheme and also present its application for a CBVCD system.

M.K. Kundu et al. (eds.), *Advanced Computing, Networking and Informatics - Volume 1*,
Smart Innovation, Systems and Technologies 27,
DOI: 10.1007/978-3-319-07353-8_36, © Springer International Publishing Switzerland 2014

The paper is organized as follows. The brief introduction is followed by the review of past work on indexing in Section 2. Details of the proposed methodology are presented in Section 3. Experimental results and concluding remarks are put in Section 4 and Section 5 respectively.

2 Past Work

A video database can be indexed following different approaches outlined in the work of Brunelli *et al.* [1]. Video data may be annotated manually at various level of abstraction. Abstraction may range from the video title to content details. The database may be organized according to the annotations to support indexing. Such manual annotation is labour intensive and subjective. To overcome the difficulties there was the development of alternate approach in the form of content-based automated indexing. A framework for automated video indexing as discussed in [1] is to segment the video data into shots and to identify the representative frames from the shots. Subsequently, the collection of those representatives is considered as the image database and experience of content based indexing of image database can be applied. Similar trend in video browsing and retrieval still persists as indicated in [2].

Zhang *et al.* [3], in their approach have classified the video shots into groups following a hierarchical partition clustering. Bertini *et al.* [4] have presented a browsing system for news video database. A temporal video management system has been presented in [5] that relies on tree based indexing. Spatio-temporal information data has been widely used for video retrieval [6,7]. Ren and Singh [8] proposed R-string representation to formulate spatio-temporal data into binary string. Non-parametric motion analysis has also been used for video indexing [9]. But, presence of noise and occlusion may lead to failure [10].

As discussed in [1], [4], the major trend for video indexing is to break it into units and to map the video database into image database by taking the representative frames of the segmented units. So it is worth to review the techniques for indexing the image database. A comprehensive study on high dimensional indexing for image retrieval has been presented in [11]. Tree based schemes are widely used. Zhou *et al.* [12] have classified those according to the indexing structure. K-D tree [13], M tree and its variants [14], TV tree [15], are few examples of such indexing structures. Hashing based techniques [16], [17] are also quite common for indexing an image database. Concept of bag of words has been deployed in number of works [18], [19].

Most of the video indexing system has focused on specific domain and thereby concentrated on designing the descriptors accordingly. Details of the underlying indexing structure have not been elaborated. It appears that the technique used for image database is adopted to cater the core need of indexing the video database.

3 Proposed Methodology

In our early work [20], a content-based video copy detection (CBVCD) system has been proposed. The system is robust enough against various photometric and post-production attacks. Each frame in the video data goes through preprocessing to reduce the effect of attacks and then features are extracted. In order to decide whether a shot in the query video is a copied version of any of the reference video shots or not, an exhaustive video sequence matching is carried out following multivariate Wald-Wolfowitz hypothesis test. Such linear search leads to large number of test which is prohibitive. In this effort, our target is to propose a methodology to reduce the number of video sequence matching.

As presented in Section 2, common approach for video indexing is to map the problem to image database indexing. The major steps are like breaking the video data into structural units called shots, extraction of representative frames of the shots to form an image database and finally well established image database indexing technique is applied on the image database formed. Proposed methodology also follows the same approach.

In this work, it is assumed that the video data is already segmented into shots following the technique presented in [21]. A clustering based shot level representative frame detection scheme is devised and subsequently indexing is done based on *triangle inequality* property [22].

3.1 Selection of Shot Representative

A shot is the collection of consecutive frames captured in a single camera session. Mostly, the frames in a shot are visually very similar. Thus, it is good enough to represent the shot by a single frame and redundancy is also removed. But, due to the motion of camera and/or object, presence of lots of activity in a shot it may not be judicious to represent the content by a single representative. Thus, multiple frames may have to be chosen for proper reflection of the content. In this context lot of works have been done [23], [24]. Here we present a simple scheme based on the following steps.

- Verify the uniformity of the shot
- For a uniform shot, select one representative frame
- For non-uniform shot, select multiple representatives based on clustering

Each frame in the shot is first represented by an edge based visual descriptor. A frame is divided into a fixed number of blocks. Normalized count of edge pixels in the blocks arranged in raster scan order forms the multi-dimensional feature vector. We have partitioned the frame into 16 blocks and same is the dimension of the vector. The features thus computed are neither too local nor global.

In order to verify the uniformity of a shot content, similarity of each frame with respect to the first one in the shot is computed. Let, s_i be the similarity value for the i-th frame in the shot. If $min\{s_i\}/max\{s_i\}$ is smaller than a threshold then the shot is taken as non-uniform one and it qualifies for multiple representatives. Similarity between two frames is measured by the Bhattacharya distance between the corresponding feature vectors and threshold is empirically determined as 0.9. For a uniform shot, frame for which the similarity with the first one is closest to $(min\{s_i\}+max\{s_i\})/2$ is taken as the representative.

For a non-uniform shot, K-means clustering is applied to put the frames into different clusters. Number of clusters, n_c is varied starting from 2 and gradually increased till optimal value for n_c is determined. Goodness of the clusters is measured based on Dunn index [25]. Once optimal numbers of clusters are formed, from each cluster the frame nearest to the centre is chosen as the representative. Thus, a shot will have number of representatives same as the number of optimal clusters.

3.2 Indexing the Database of Representative Frames

Indexing scheme is applied on the image database obtained after collecting the representative frames of all the shots in the video sequences. Based on the *triangle inequality* approach, a value for each database image is assigned which corresponds to the lower bound on the distance between database image and a query image. On that value, a threshold is applied to discard the database images which are away from the query image.

Let $I = \{i_1, i_2, . . . , i_n\}$ and $K = \{k_1, k_2, . . . , k_m\}$ denote the database of images and collection of key images chosen from I respectively. As per *triangle inequality,* $dist(i_p,Q) + dist(Q, k_j) \geq dist(i_p, k_j)$ is true where $i_p \in I, k_j \in K$, Q is a query image and $dist(..)$ stands for a distance measure. The equation can be rewritten as $dist(i_p,Q) \geq | dist(i_p, k_j) - dist(Q, k_j) |$. Considering all $k_j \in K$, the lower bound on dist(i_p,Q) can be obtained from $dist(i_p,Q) \geq max_j(| dist(i_p, k_j) - dist(Q, k_j) |)$ where $j \in \{1, 2, . . . ,m\}$. Thus, the major steps are as follows.

- Select the key images K from the image database I
- pre-compute the distance matrix to store $dist(i_p, k_j)$ for all $i_p \in I$ w.r.t all $k_j \in K$

All these steps are offline. In order to select the key images, images in the database are partitioned into clusters following k-means clustering algorithm and number of optimal clusters is decided based on Dunn-index. For each cluster, image nearest to the cluster centre is taken as key image. In order to prepare the distance matrix, n ×m of computation is required. In our experiment, $dist(i_p, k_j) = 1 - bhatt\_dist(i_p, k_j)$ where $bhatt\_dist(i_p, k_j)$ provides the similarity between i_p and k_j .

For a CBVCD system, our point of interest is to find out the images from the database for which the lower bound does not exceed a threshold t. For searching against a query image(Q) the steps to be carried out in online mode are as follows.

- Compute $dist(Q, k_j)$ for all $k_j \in K$
- Retrieve all i_p such that $| dist(i_p, k_j) - dist(Q, k_j) | \leq t$ *for all* k_j *in* K

The first step requires m number of computation whereas the second step involves n × m simple operations if searched linearly. The comparison can be reduced further based on the actual implementation. In our work, with respect to each k_j a separate structure stores $dist(i_p, k_j)$ in an order. It enables binary search to retrieve the i_ps satisfying the condition $dist(Q, k_j) - t \leq dist(i_p, k_j) \leq dist(Q, k_j) + t$. Thus, the overhead of accessing the index is also reduced significantly. In our experiment value of t is empirically determined and taken as 0.01.

Now, for video copy detection, representative frames are extracted from the shot to be tested. Each such representative frame is used as the query image (Q) to retrieve the desired images from the database. With the shots corresponding to all the retrieved images, the final hypothesis test is carried out to decide the outcome. Thus the final test is carried out with only a small subset of the shots in the reference database making the CBVCD system faster.

4 Experimental Result

In order to carry out the experiment we have worked with a collection of video data taken mostly from TRECVID 2001 dataset and a few other recordings of news and sports program. The sequences are segmented into shots following the methodology presented in [21]. The dataset contains 560 shots. Performances of the proposed methodology to extract the representative frames of the shot and effectiveness of indexing are studied through experiments.

All the shots are manually ground-truthed and marked as single or multiple depending on whether a shot requires a single frame or more than one frame as its representative. For shots of *multiple* types, different homogeneous sub-shots are also noted. Performance of the proposed scheme for extracting the representative frames is shown in Table 1. It shows that the homogeneous shots are correctly identified as single and it fails only for a few cases of shots of type *multiple*. There may be other type of errors like *over-splitting* and *under-splitting* in case the number of extracted frame is more or less than the expected count respectively. In our experiment, only five shots are under-splitted extracted frames is one less than the desired count and over-splitting occurred for three shots.

To study the performance of the indexing system, we have considered the application of CBVCD. Index based search has been incorporated in the CBVCD system presented in [20]. In linear method, a query video shot is verified with all the shots in the reference video database using hypothesis test based sequence matching technique. But in case of indexed method, following the proposed scheme a subset of shots whose representative frames match with those of query shot are retrieved. Only with the retrieved shots final verification is made. Adoption of indexing scheme is expected to make the process faster but it may affect the detection performance. The performance of a CBVCD can be measured in terms of correct recognition rate (CR) and false alarm rate (FR). These are measured as $CR=(n_c/\ n_a)$ x $100\ \%$ and $FR = (n_f/n_q)$x$100\%$where n_c, n_a, n_f , n_q are number of correctly detected copies, number of actual copies, number of false alarm and number of queries respectively.

Table 1. Performance of Extraction of Representative Frame from Shot

Shot type	Number of shots	Detected as	
		Single	Multiple
Single	489	489	0
Multiple	71	4	67

Table 2. Performance of CBVCD system

Attack type	No. of query	Linear method		Indexed method	
		CR(in%)	FR(in%)	CR(in%)	FR(in%)
No attack	560	100.00	18.57	100.00	15.18
Brightness change	300	100.00	23.67	99.67	11.33
Contrast change	300	100.00	22.67	99.67	11.00
Noise Corruption	200	99.00	19.00	98.50	10.50
Flat file	100	99.00	18.00	98.00	13.00
Letter box	100	98.00	16.00	98.00	14.00
Pillar	100	98.00	16.00	97.00	14.00
Logo insertion	100	98.00	18.00	94.00	15.00
Overal	1760	99.49	19.82	98.98	8.47

Table 2 shows the performance of linear and indexed method for the CBVCD system. In case of indexed method the sequence matching is carried out in a reduced search space. As a result, exclusion of similar sequence is possible, particularly for the transformed version. But, for a CBVCD system similarity is not synonymous with copy. It is well reflected in Table 2 that reduction of search space has reduced both CR and FR. But, CR is reduced marginally and FR (it is more sensitive and should have low value) is reduced significantly. Even under different photometric (change in brightness, contrast, corruption by noise) and post-production (insertion of logo, change in display format – flat file, letter box, pillar) attacks, the indexed method reduces FR substantially without much compromising the correct recognition rate.

As the indexing scheme reduces the search space, the copy detection process as a whole becomes faster. Experiment with 1760 query shots has revealed that on an average the system becomes five times faster.

5 Conclusion

In this work a simple but novel video indexing scheme is presented. A clustering based scheme is proposed which dynamically determines the number of representative frames and extracts them. A triangle inequality based indexing scheme is adopted for the image database formed by collecting the representative frames for all the shots. For a shot given as the query, candidate shots are retrieved based on the proposed methodology. On the retrieved candidate shots, video sequence matching technique can be applied to fulfill the requirement of a CBVCD system. Experiment indicates the effectiveness of the proposed methodology in extracting the representative frames. Applicability of the indexing scheme in CBVCD system is also well established as it reduces false alarm rate drastically without making much compromise on correct recognition rate and it speeds up the process significantly.

References

1. Brunelli, R., Mich, O., Moden, C.M.: A survey on the automatic indexing of video data. Journal of Visual Communication and Image Representation 10, 78–112 (1999)
2. Smeaton, A.F.: Techniques used and open challenges to the analysis, indexing and retrieval of digital video. Information Systems 32, 545–559 (2007)
3. Zhang, H.J., Wu, J., Zhong, D., Smoliar, S.W.: An integrated system for content based video retrieval and browsing. Pattern Recognition 30(4), 643–658 (1997)
4. Bertini, M., Bimbo, A.D., Pala, P.: Indexing for reuse of tv news shot. Pattern Recognition 35, 581–591 (2002)
5. Li, J.Z., OZsu, M.T., Szafron, D.: Modeling video temporal relationships in an object database systems. In: Proc. SPIE Multimedia Computing and Networking, pp. 80–91 (1997)
6. Pingali, G., Opalach, A., Jean, Y., Carlbom, I.: Instantly indexed multimedia databases of real world events. IEEE Trans. on Multimedia 4(2), 269–282 (2002)
7. Ren, W., Singh, S., Singh, M., Zhu, Y.S.: State-of-the on spatio-temporal information-based video retrieval. Pattern Recognition 42 (2009)
8. Ren, W., Singh, S.: Video sequence matching with spatio-temporal constraint. In: Intl. Conf. Pattern Recog., pp. 834–837 (2004)
9. Fablet, R., Bouthmey, P.: Motion recognition using spatio-temporal random walks in sequence of 2d motion-related measurements. In: Proc. Intl. Conf. on Image Processing, pp. 652–655 (2001)
10. Fleuret, F., Berclaz, J., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. IEEE Trans. on PAMI 20(2), 267–282 (2008)
11. Fu Ai, L., Qing Yu, J., Feng He, Y., Guan, T.: High-dimensional indexing technologies for large scale content-based image retrieval: A review. Journal of Zhejiang University-SCIENCE C (Computers & Electronics) 14(7), 505–520 (2013)
12. Zhou, L.: Research on local features aggregating and indexing algorithm in large-scale image retrieval. Master Thesis, Huazhong University of Science and Technology, China 10–15 (2011)
13. Robinson, T.J.: The k-d-b tree: A search structure for large multidimensional dynamic indexes. In: Proc. ACM SIGMOD Intl. Conf. on Management of Data, pp. 10–18 (1981)
14. Skopal, T., Lokoc, J.: New dynamic construction techniques for m-tree. Journal of Discrete Algorithm 7(1), 62–77 (2009)
15. Lin, K.I., Jagadish, H.V., Faloutsos, C.: The tv-tree: An index structure for high-dimensional data. VLDB Journal 3(4), 517–542 (1994)

16. Zhuang, Y., Liu, Y., Wu, F., Zhang, Y., Shao, J.: Hypergraph spectral hashing for similarity search of social image. In: Proc. ACM Int. Conf. on Multimedia, pp. 1457–1460 (2011)
17. Heo, J.P., Lee, Y., He, J., Chang, S.F., Yoon, S.E.: Spherical hashing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2957–2964 (2012)
18. Avrithis, Y., Kalantidis, Y.: Approximate gaussian mixtures for large scale vocabularies. In: Proc. European Conf. on Computer Vision, pp. 15–28 (2012)
19. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Trans. PAMI 33(1), 117–128 (2011)
20. Dutta, D., Saha, S.K., Chanda, B.: An attack invariant scheme for content-based video copy detection. Signal Image and Video Processing 7(4), 665–677 (2013)
21. Mohanta, P.P., Saha, S.K., Chanda, B.: A model-based shot boundary detection technique using frame transition parameters. IEEE Trans. on Multimedia 14(1), 223–233 (2012)
22. Berman, A.P., Shapiro, L.G.: A flexible image database system for content-based retrieval. Computer Vision and Image Understanding 75(1/2), 175–195 (1999)
23. Ciocca, G., Schettini, R.: An innovative algorithm for key frame extraction in video summarization. Real Time IP 1, 69–98 (2006)
24. Mohanta, P.P., Saha, S.K., Chanda, B.: A novel technique for size constrained video storyboard generation using statistical run test and spanning tree. Int. J. Image Graphics 13(1) (2013)
25. Dunn, J.C.: Well separated clusters and optimal fuzzy partitions. Journal of Cybernetica 4, 95–104 (1974)