THE UNIVERSITY OF SHEFFIELD AND UNIVERSITY OF ENGINEERING & TECHNOLOGY, LAHORE AT TECVID 2014: INSTANCE SEARCH TASK

Sana Amanat[†]*,Muhammad Usman Ghani Khan*[†]*, Nudrat Nida*[†] *and Yoshihiko Gotoh*[‡]

[†] Department of Computer Science & Engineering,
UET Lahore Pakistan.
[‡] Department of Computer Science,
University of Sheffield, UK

ABSTRACT

In this paper, we aim at presenting our contribution related to Instance Search(INS) tasks of TRECVID 2014. For instance search our approach is based on SIFT and Covariant Feature Detector, second approach computes HOG and Distance Transformation, third approach is a combination of distance transform only. In fourth approach we presents a visual information retrieval system by computing relevance to a given query image. This approaches is based on Content-based Image Retrieval, using global features color & texture characteristics. Feature vectors are combination of color and edge directivity descriptor CEDD [1].

Index Terms— video retrieval, instance search task, video indexing ,Content-based Image Retrieval (CBIR).

1. INTRODUCTION

In 2010 TRECVID campaign introduce INS task to promote progress in content-based analysis of and retrieval from digital video via open, metrics-based evaluation. In TRECVID 2014[2];the testing data was produced from the BBC easter elder dataset collection. We segmented the video clips into master shots that semantically describes the scenario. Then using the time elements we extracted the frame of video. Using shot boundary detection key frame is generated, to reduce the duplication and decrease the complexity of image retrieval process. There were a 243 test video clips and 29 image test queries. Some image transformations were also applied to random test clips. The task includes recurring queries with people, location and objects in the rushes.

This year, there were 29 topics and more than 70000 short clips as testing data collected from the BBC East Elder Dataset ;"Programme material copyrighted by BBC". The main objectives from our participant was to explore the task definition and the evaluation issues.

2. INSTANCE SEARCH TASK

For INS task we submitted four runs. Following sections present detailed discussion of these runs.

2.1. Run 1: SIFT and Covariant Feature Detector Features

2.1.1. Framework Overview

The whole framework is shown in Figure 1. The first step is to segment the video into pieces, since in this year, the video is given as an original form, and some of them last close to 2 hours. Then for each segment, the key frame is extracted and the further searching is based on key frame only for our calculation ability. For each segment, only one key frame is extracted. During the searching stage, since we want to combine advantages of both global and local descriptors. Then we normalized the score for each feature and fusion them together.

2.1.2. Segment stage

The video this year seems like a movie or TV play with voice and coherent plot. For some cases, it is not easy to detect the boundary, especially for the scene with non abrupt change. In order to catch each boundary changing on the video content, here, we adopt the shot boundary detection using X-Or differences and extract the key frame from each shot. This year, the segment label is given in a list. So in our segmentation stage, we relax selection to make sure the non-segments will be combined together. So our segments number is much more than the labeled one. Compared with the given label list, we re-label our segments, and maybe several continuous segments belongs to the same label. Furthermore, we extract the middle frame as the key frame in our segment list, and each segment is with only one key frame for the calculation ability.

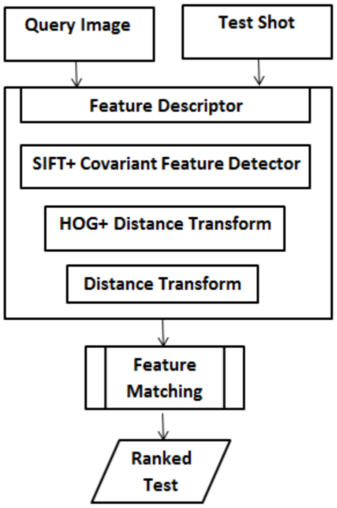

Fig. 1. Framework of searching for run 1 ,2 & 3

2.1.3. Matching stage

In this stage, we combined local descriptor and global one together, since for some query image, the target object is too small to get enough information by local descriptor (shown in 1). For both key frame sequence and the query image after background subtraction,then extracting SIFT and Covariant Feature Detector for a query image and all test shots. Each SIFT key-points in the query topic is matched to its corresponding descriptors in the video clip database as proposed in [3]. After that these computed features of a query image is matched with the features of key frames of all test shots one by one. This matching is performed using Euclidean distance between query and test shot. Higher value of Euclidean distance depicts the high rate of dissimilarity. All these calculated distances are sorted in ascending order to get ranking of videos with respect to similarity rate. This procedure is performed for all query images against provided test shots.

2.2. Run 2: Baseline run with HOG descriptors and Distance Transformation Features

Similar to the first run, one frame per second are extracted from every video clips and using segmentation algorithm to extract the key frame from each shot, we reduces the number of duplicated frames. HOG features are extracted for a query image and all test shots are calculated. Based on extracted HOG; Distance transform DT descriptor is calculated using euclidean distances between query image and all test shots. Smaller the Euclidean distance higher the similarity rate; all test shots are sorted based on the distance from query image and test shots. We ranked the images by query image based on difference between the query image and key frame. Smaller the difference higher rank is assigned to key frames.

2.3. Run 3:Run with Distance Transformation Features

The framework of run 3 with distance transform feature searching is presented in part of Figure 1. Given the image set of topic, we extracted the Region of Interest (ROI) using the related mask. Then the feature vector consists of distance transform are computed. For the search. The computed scores based on the squared Euclidean distance between the query topic descriptor and the closest descriptor in the video database. Finally, the highest scores are used as rank in the final result. Evaluation results for this run are presented in figure 1

2.4. Run 4: IR based on Visual Semantics

An IR-based framework is proposed to efficiently retrieve candidate images from large source collections. The source collection is indexed off line. The testing image is split into smaller queries. The index is queried against each query from the testing image to retrieve a set of potential source video segments. The top N images are selected for each testing image and the results of multiple queries merged using a score-based fusion approach [4] to generate a ranked list of source videos. The top K images in the ranked list generated by CombSUM are marked as potential candidate images.

Figure 2 shows the proposed process for retrieving candidate images using an IR-based approach. The source collection is indexed with an IR system (an offline step). The candidate retrieval process can be divided into four main steps: (1) pre-processing, (2) query formulation, (3) retrieval and (4) result merging. These steps are described as follows:

1. **Pre-processing:** This is the step for feature generation. Similar to the first two runs, for each of the suspicious document, MPEG-7[5], CEDD (colour edge directivity descriptor) and FCTH(Fuzzy Color and Texture Histogram)[6] features are calculated and histograms of those features are generated. These feature values are used as index of key frame to represent that particular frame.

2. **Query Formulation:** Similarly features from query image is calculated. These calculated features are used for comparison between the query image and key-frames.

3. **Result Merging:** The top N source documents from the result sets returned against multiple queries are merged to generate a final ranked list of source documents. In a list of source documents retrieved from a query, document(s) at the top of the list are likely to be the similar videos. In addition, portions of text from a single source document can be reused at different places in the same video segment. Therefore, selecting only the top N documents for each query in the result merging process is likely to lead to the original source

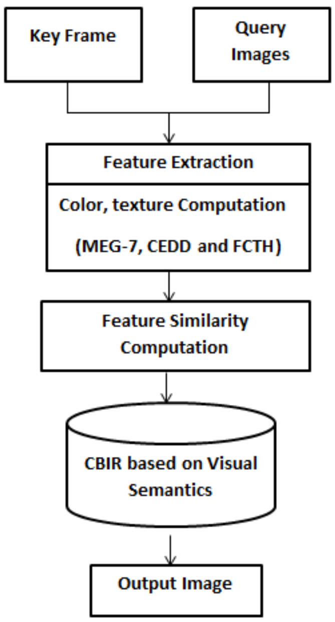
Key Frame
Query Images
Feature Extraction
Color, texture Computation
(MEG-7, CEDD and FCTH)
Feature Similarity Computation
CBIR based on Visual Semantics
Output Image

Fig. 2. Process of candidate document retrieval

document(s) appearing at the top of the final ranked list of the documents.

A standard data fusion approach called CombSUM method [4] is used to generate the final ranked list of documents by combining the similarity scores of source documents retrieved against multiple queries. In the CombSUM method, the final similarity score, $S_{finalscore}$, is obtained by adding the similarity scores of source documents obtained from each query q:

$$S_{finalscore} = \sum_{q=1}^{N_q} S_q\left(d\right) \tag{1}$$

where N_q is the total number of queries to be combined and $S_q\left(d\right)$ is the similarity score of a source document *d* for a query q.

The top K documents in the ranked list generated by the CombSUM method are marked as potential candidate source documents.

2.4.1. Implementation

One of the popular and freely available Information Retrieval systems are used to implement the proposed IR-based framework:based on (1) LIRE [7]. In both Terrier [8] and Lucene [9] Terrier and Lucene, terms are weighted using the *tf.idf* weighting scheme. In Terrier, documents against a query term are matched using the TAAT (Term-At-A-Time) approach, each query term is matched against all posting lists to compute the similarity score. In Lucene, the similarity score between query and document vectors is computed using the cosine similarity measure. We have used LIRE that retrieve images based on the visual semantics of the image. Using key frames an indexer is created and some of the low level features are computed against each query image and all test shots such as MPEG-7, CEDD and FCTH. Based on the similarity ratio of features shots are indexed and top 993 shots are selected as an output. ....

3. CONCLUSION

In this paper we presented our experiments performed in the TRECVID 2014 instance search tasks. This participation rewarded us an experience in our researches and in finding new ideas and directions in the domain of object-based video retrieval.

4. REFERENCES

[1] Savvas A Chatzichristofis and Yiannis S Boutalis, “Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval,” in *Computer Vision Systems*, pp. 312–322. Springer, 2008.

[2] Paul Over and George Awad and Martial Michel and Jonathan Fiscus and Greg Sanders and Wessel Kraaij and Alan F. Smeaton and Georges Quenot, “ TRECVID 2014 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics,” in *Proceedings of TRECVID 2014*. NIST, USA, 2014.

[3] David G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, pp. 91–110, 2004.

[4] E. Fox and J. Shaw, “Combination of multiple searches,” *NIST SPECIAL PUBLICATION SP*, pp. 243–243, 1994.

[5] Thomas Sikora, “The mpeg-7 visual standard for content description-an overview,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 11, no. 6, pp. 696–702, 2001.

[6] Savvas A Chatzichristofis and Yiannis S Boutalis, “Fcth: Fuzzy color and texture histogram-a low level feature for accurate image retrieval,” in *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS’08. Ninth International Workshop on*. IEEE, 2008, pp. 191–196.

[7] Lire Ersado, “Child labor and schooling decisions in urban and rural areas: comparative evidence from nepal, peru, and zimbabwe,” *World Development*, vol. 33, no. 3, pp. 455–480, 2005.

[8] I. Ounis, G. Amati, Plachouras V., B. He, C. Macdonald, and Johnson, “Terrier Information Retrieval Platform,” in *Proceedings of the 27th European Conference on Information Retrieval*. 2005, pp. 517–519, Springer.

[9] E. Hatcher, O. Gospodnetic, and M. McCandless, *Lucene in Action*, Manning Publications, 2004.