RMIT University at TRECVID 2011 Content-Based Copy Detection

Amir H. Rouhi
Computer Science and Information Technology
RMIT University
Melbourne, Australia
Email: amir.rouhi@rmit.edu.au

James A. Thom
Computer Science and Information Technology
RMIT University
Melbourne, Australia
Email: james.thom@rmit.edu.au

***Abstract*—In this paper, we introduce our approach used for TRECVID 2011 *Content-Based Copy Detection* (CCD) task. It was the first experience of RMIT University in TRECVID and with respect to the team background in image processing and using global features in this field of research we preferred to follow the same technique in video copy detection task in this year. Our presented method is based on extracting frames from video stream. Before extracting features from the images, we perform some preprocessing phases on the extracted images to reduce the complexity. To build the multidimensional feature vector structure, global features such as *Auto-Correlogram* and *Intensity Histogram* are used on two different regions of the images. The same method is used in four different runs submitted by RMIT and each run was varied with other runs in their threshold (V) value.**

I. Introduction

There are two major approaches in video only CCD projects. The first approach, which is the approach we use, is based on Content Based Image Retrieval (CBIR) techniques. It is usually referred as Frame Based video retrieval. In this approach, some key frames or any frames with fixed interval distance will be extracted from the video stream in the form of colored images. Then based on the known techniques in CBIR, some features will be extracted from the selected images. The features selected can include local features or global features or both kind of features [5]. Local features are mostly based on textures in regional areas of an image [4] but global features represent color presentation and shape description [4] on the whole area on which they applied. In the retrieval phase, the distances between the selected images from query video and all the database videos are calculated and the minimum distance (or maximum similarity) on every pair segment of videos, represent the copy. The second major approach is totally different and is based on motion vector. The similarity of motion objects in consecutive frames represents the similarity of segments of videos.

We used global features in two different regions of images in our proposed feature vector structure. Generally, global features are known as not robust features against clutter and occlusions. The results shows that in spite of using just an algorithm in different runs and just emphasizing on global features only, the algorithm is able to detect some highly transformed video copies mostly in:

- T3: Insertions of pattern
- T4: Compression
- T6: Decrease in quality
- T8: Post production
- T10: change to randomly choose

It was due to the preprocessing phases and regionalizing the image and the combination of selected global features in proposed algorithm. Though the overall score of TRECVID shows that all of these results were under the median line but we noticed the algorithm has performed very good detection in some severely transformed videos. One reason for the overall poor score is that we did not arrange a suitable result set of our algorithm. We tried to remove the redundancy for algorithm result set and finally submitted only three selected answers for each query. The second thing that our algorithm suffered from is that it works just based on extracting visual features and not any audio features.

Conversely the mean processing time of the proposed algorithm shows acceptable results, slightly better than median processing time for most of the transformations.

II. Video Only Content Based Copy Detection

A. Indexing Phase

The flowchart of proposed indexing algorithm is presented in Figure 1. As can be seen in the flow chart, the algorithm for both groups of videos, query and referenced videos offered a similar approach. The approach is generally based on frame sampling from both types of videos in fixed rate of one second intervals of the video stream.

We designed the algorithm based on the assumption that, the most important part of images is intensity plane of the image. Y-plane or intensity information of image pixels, represents the content of the image and describes the shapes of the image effectively. For this, we use the open source software, ffmpeg in our C++ code and save extracted Y plane images in an appropriate folder for performing the next stages of the algorithm. We reduces the complexity of CCD task by using only Y plane from the three image planes: Y, U and V. The number of videos in TRECVID 2011 was about 8300 referenced video and 1600 query videos in range of 10

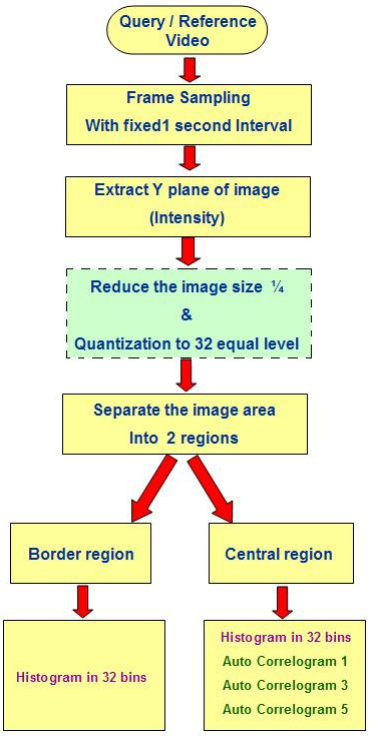

Fig. 1. indexing phases.

second to 3.5 minutes [2]. If we consider the average video length as 100 seconds, the overall extracted images would be 830,000 images of referenced videos and 160,000 images of query videos. In search algorithm we searched all query images against all referenced video images. It will come up with approximately 132 billion image comparison which force a huge computational time on search algorithm. Due to such huge number of comparison in searching algorithm, we decided to reduce the complexity of the problem in another stage. This stage can be seen in the dashed box of the flowchart shown in Figure 1. In this stage, the size of the intensity images reduced to 1/4th of the original size and the intensity levels quantized to 32 equal bins. Each bin contains 8 intensity levels from 0 to 255. After these three stage of complexity reduction, the algorithm starts to extract a combination of global features in two regions of each image.

According to a basic rule of composition in photography, the most important points of an image are four points by dividing the image to thirds, both in horizontal and vertical [1]. With respect to this basic rule, for capturing more important object located in golden ratio points, we divide the image to border region and central region. The border area covers the outside area by dividing the image in sixth in horizontal and vertical. Figure 2 shows these two regions and the four important points.

The indexing algorithm provides two categories of features for the two regions of images. With respect to less importance

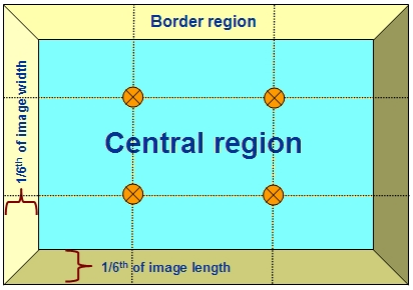

Fig. 2. Four important point of an image according to the rule of thirds or golden ratio in photography shown by circles on intersections. The Border region and central region are shown in this figure too.

of the border area, we use the Histogram of the quantized intensity values in 32 bins. But with respect to importance of the central area, we use a combination of Histogram and Auto-Correlogram [3]. These two types of global features are implemented on the quantized intensity values of pixels located in central area on 32 bins. Auto-Correlogram has been computed for pairs of pixels in three distances: one, three and five pixel distance. The three different distances on pairs of same colored pixels offers an acceptable scale invariant feature. These features in combine with Histogram represents an effective global feature structure. One of the deficiencies of Correlogram is its high computational time which is $O(n^2d)$ for the number of distances d.
We used Auto-Correlogram on pairs of same colored pixels which has the size of $O(nd)$ in time. The final structure of the feature vector of every single image contains five array of 32 real numbers which correspond to:

- Histogram of border area.
- Histogram of central area.
- Auto-Correlogram in 1 pixel distance of central area.
- Auto-Correlogram in 3 pixel distance of central area.
- Auto-Correlogram in 5 pixel distance of central area.

B. Searching phase and sequence matching

The output of the index algorithm is indexed videos which are the input of searching algorithm as can be seen in Figure 3. The searching algorithm computes the distances of all images in query videos against all the images of all the referenced videos. The result of each internal loop of searching algorithm is a three dimensional [I x J x 5] distance matrix (I represents the number of retrieved images in query video, J represents the number of retrieved images in referenced video and 5 represents the five dimension of feature vector structure)which each of its elements represents the distance of i^{th} image of query video to the j^{th} image of referenced video images in each of the five series elements of features (Figure 4).

The proposed distance method for searching algorithm is based on Manhattan distance. Manhattan distance will be computed on each of the five types of features and finally these five parts will be summed up with appropriate weight

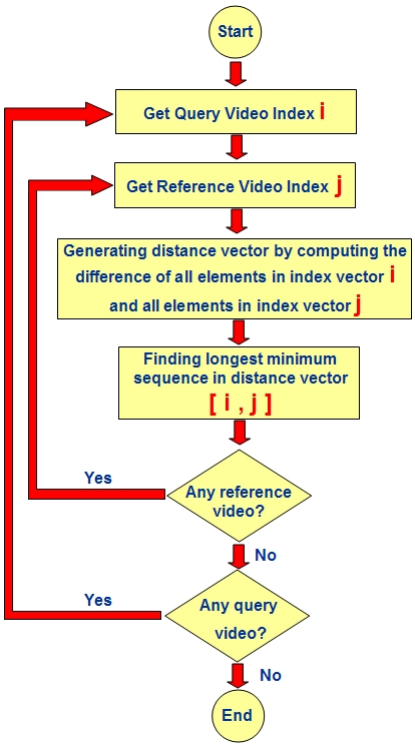

Fig. 3. General flowchart of the searching algorithm.

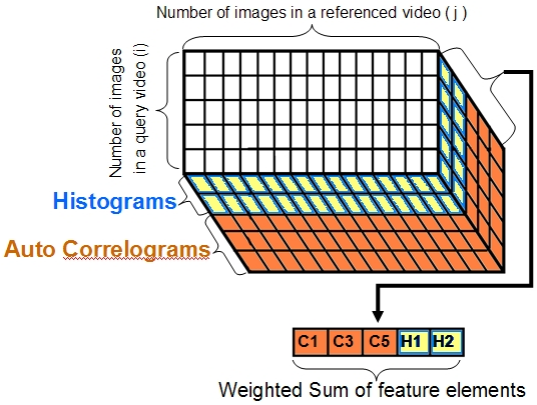

Fig. 4. First three dimensional distance matrix .

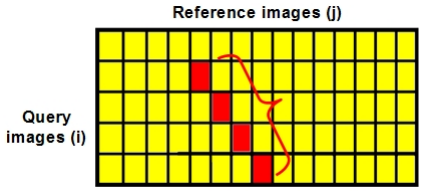

Fig. 5. Final two dimensional distance matrix. The longest minimum diagonal sequence (highlighted boxes) shows the matching part of query video i and reference video j.

for each type of feature. As the central area features are more important than the border area, the weight of central area features are bigger than border area. The three types of central area features have different weights as well. The result of weighted sum of the features is a two dimensional [I x J] distance matrix(Figure 5).

The most important and crucial part of the searching algorithm is the alignment method used for finding the longest minimum diagonal sequence in the distance matrix$[i, j]$. The size of this two dimensional matrix is equal to the number of extracted images of query video as row and number of extracted images of each reference video as column. If there exist a matching segment or copy of query video i in referenced video j, the sum of values of diagonal elements of the matrix shows the minimum values of difference in comparison with all other diagonal summations. Finding the longest diagonal sequence with minimum sum of distance values needs a robust method in this part of algorithm. The length of the minimum diagonal sequence represents the time length of the copied video in seconds. The matching start time in the query can be investigated by first index number of rows. The same is for matching start time for reference video. It is related to the first index number of the matched columns. As we extracted the images from video stream in one second intervals, the searching algorithm can just represent start and end time of matching parts by integer numbers. Figures 4 and 5 tries to illustrate this stage of the algorithm.

C. Results

Our approach was based on investigating the similarity detection among randomly extracted images of randomly selected videos. The graph shown in Figure 6 demonstrates this similarity. After selecting a random image, e.g. the 34^{th} image of the reference video: 2-21-2005food-couscous. _-o-_.food_couscous_ 512kb .mp4, the result of the proposed algorithm in finding the most similar frames to the 34^{th} frame is shown in this graph. The weighted sum difference of the 34^{th} image features with all the other images is shown by the dashed top line on the graph. Naturally the minimum difference value is zero and belongs to the 34^{th} image with itself. The frames with minimum differences can be seen in the solid rectangle on the graph. The orders of similar frames are shown by big numbers in Figure 7. The graph in Figure 6 shows how important the weighted sum of differences of the five features (dashed line) is to discriminate the image similarity in comparison with other global features.

1) TRECVID Results: Although we only submitted 3 matches for each query video, analyzing the TRECVID results shows that the algorithm is performing acceptable results in some types of transformations as mentioned in introduction. Some of the results of the algorithm can be seen in the Table 1.

The submitted runs were consist of two Balanced runs and two NOFA runs as following:

1) RMIT.m.balanced.VideoBal5 $(V = 0.5)$
2) RMIT.m.balanced.VideoBal6 $(V = 0.6)$

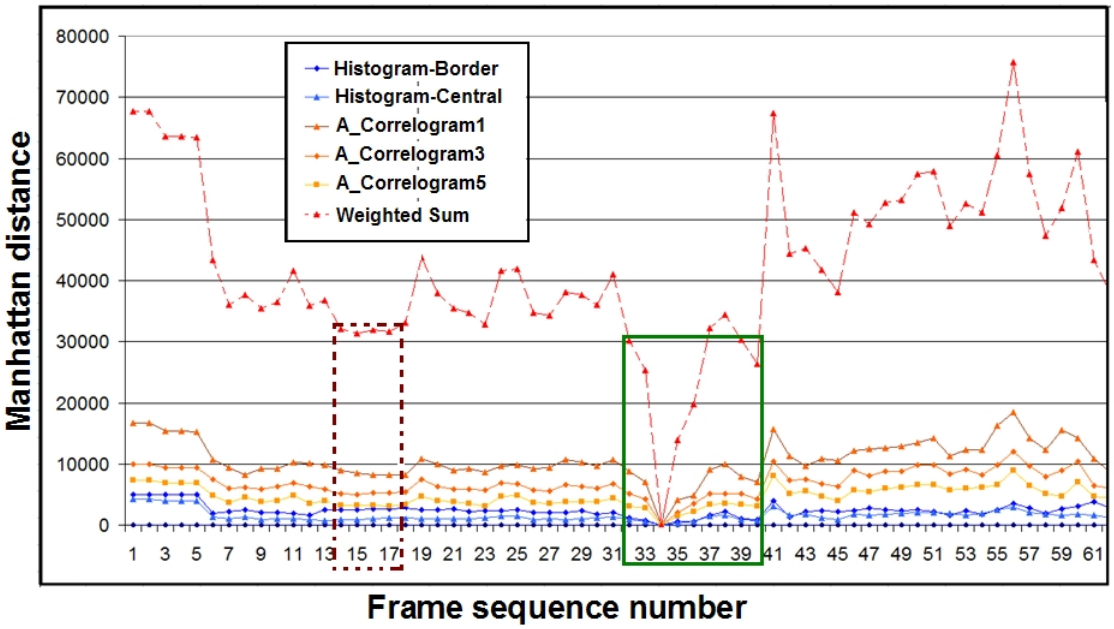

Fig. 6. The frame similarity graph shows the 34^{th} frame and its most similar frames in solid line rectangle with distance threshold: 30,000. The dashed rectangle shows the probable false alarm area if the distance threshold value set to a value more than 30,000.

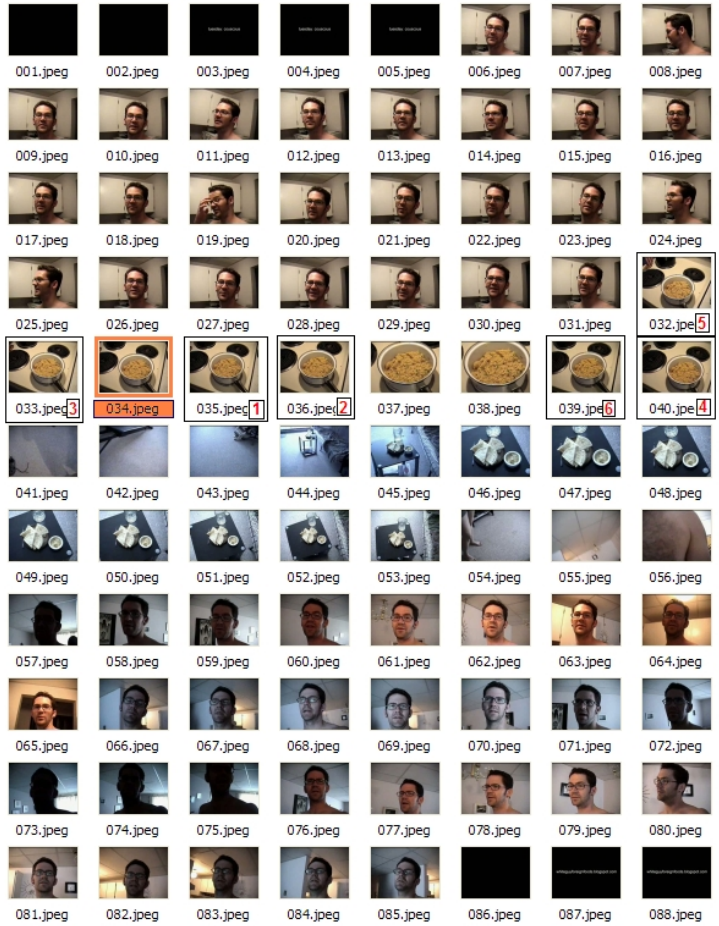

Fig. 7. Sequential frames extracted from a TRECVID video with 1 second interval. The top 6 similar images are captured in black rectangles. The numbers on down right corners represents their similarity order to the randomly selected frame no: 34.

Ref. Video id.	Query Video id.	First ref. Video Time	Last ref. Video Time	First Query Time	Perc.
8882	9786,7054,6984,8536	67	99	0	0.99
8112	12730,6509,5239	39	90	0	1
5700	6148,7729	41	93	1	1
5700	6686	125	134	0	0.96
6524	6797	78	109	0	0.99
6524	5926	103	109	25	1
8173	6563,9162,9405,5975	66	117	0	1
5515	8041,8133	65	107	1	1
11064	7331,8661	49	80	0	1
11064	9986	56	79	7	1

TABLE I
SOME CCD RESULTS OF THE PROPOSED ALGORITHM.

3) RMIT.m.nofa.VideoNOFA7 $(V = 0.7)$
4) RMIT.m.nofa.VideoNOFA8 $(V = 0.8)$

As described earlier, the content of all four runs are exactly same results of the same algorithm (due to lack of enough time to submit really different results consequent of the changes in the algorithm). The only difference among submitted runs were the threshold value V. This parameter can be seen at the end of each runs name. The results of actual Balance with V=0.5 can be seen on the Figure 8. Two DET(Detection Error Trade-off) plot of transformations No:16 and No:23 can be seen in Figures 9 and 10 respectively.

III. CONCLUSION

As can be seen in the following TRECVID graphs, the overall result is not very good mostly because we only submitted at most three matches for each query video. The mean processing time is similar to other algorithms.In regard with TRECVID evaluation analysis and our experiments, our general conclusion regarding using global features for CCD or any content based video retrieval is that global features are effective for finding matching sequences in video streams but they suffer from a high false alarm rate (RFA) as shown in Figure 6 (dashed rectangle). For acquiring the best results in frame based CCD, a combination of global features and local scale invariance features is necessary. In that case the trade-off between efficiency and effectiveness would be another challenge for our research.

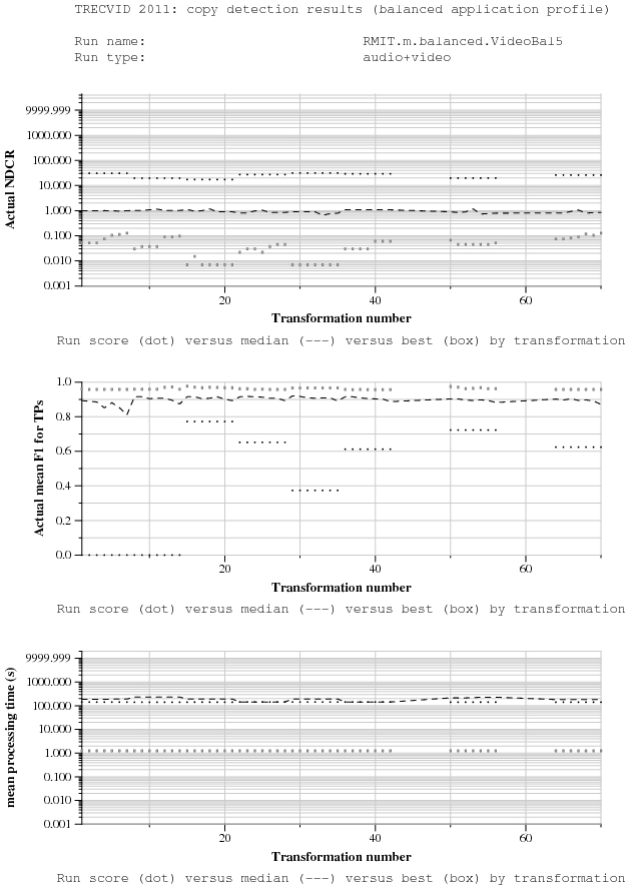

Fig. 8. Statistical graphs of actual balanced with threshold 0.5.

REFERENCES

[1] http://en.wikipedia.org/wiki/Rule-of thirds. Wikipedia.
[2] IACC.1A http://www-nlpir.nist.gov/projects/tv2011/tv2011.html IACC.1.A. *TRECVID 2011 web page*, 2011.
[3] J. Huang, S.R. Kumar, M. Mitra, W.J. Zhu, and R. Zabih. Image indexing using color correlograms. In *CVPR*, page 762. Published by the IEEE Computer Society, 1997.
[4] D.A. Lisin, M.A. Mattar, M.B. Blaschko, E.G. Learned-Miller, and M.C. Benfield. Combining local and global image features for object class recognition. In *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 43–47. IEEE, 2005.
[5] Y. Uchida, S. Sakazawa, M. Agrawal, and M. Akbacak. *KDDI Labs and SRI International at TRECVID 2010: Content-based copy detection*, 2010.

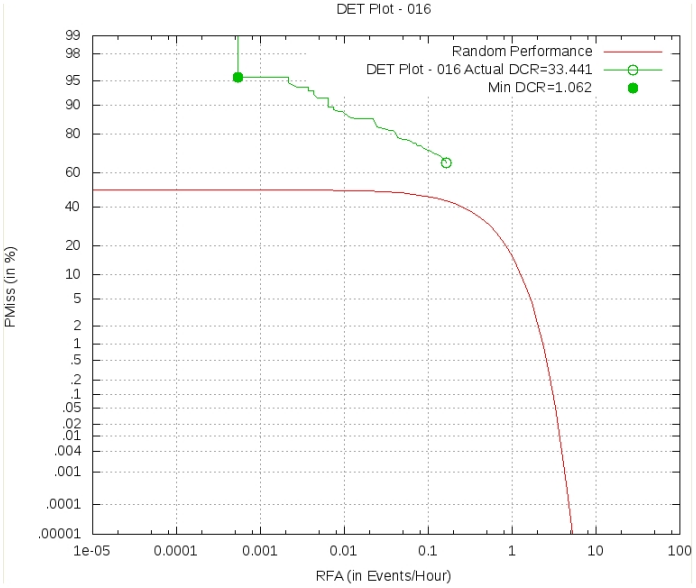

Fig. 9. DET plot of transformation No:16.

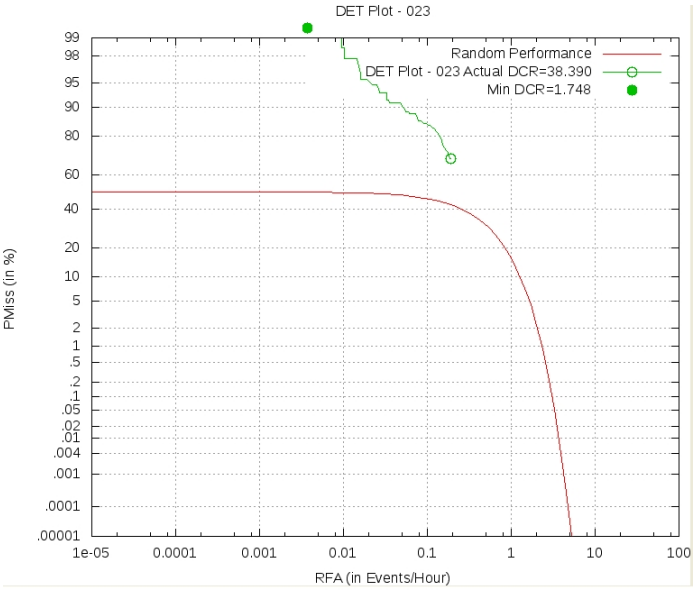

Fig. 10. DET plot of transformation No:23.