Glasgow University at TRECVid 2007

P. Punitha
Department of Computing Science
University of Glasgow, UK
punitha@dcs.gla.ac.uk

Joemon M Jose
Department of Computing Science
University of Glasgow, UK
jj@dcs.gla.ac.uk

Abstract

In this paper we describe our experiments in the automatic search task of TRECVid 2007. For this we have implemented a new video search technique based on SIFT features and manual annotation. We submitted two runs, one solely based on the SIFT features with keyframe matching and the other based on adapted SIFT features for video retrieval in addition to manually annotated data.

1. Introduction

This year Glasgow University participated in summarisation and automatic search task, whereas in the previous year, automatic and interactive search results were submitted. This year we submitted two fully automatic runs. Amongst the automatic runs, one (UG_F_Sys1) is based on matching SIFT features and the other **(**UG_F_Sys2**)** based on adapted SIFT features and annotated data. The following describes the submitted runs:

UG_F_Sys1: Automatic search based on SIFT features

UG_F_Sys2: Automatic search based on text and SIFT features

Both of our runs were of type c, and no other data provided were used for training. Both runs were trained on the TRECVid 2007 development set only.

The remainder of the paper is organised as follows. In section 2, we describe the features used. The details of the submitted runs are given in section 3. Section 4 discusses the results and the paper concludes in section 5.

2. Features

The basic unit of processing is the video shots. We used the common shot boundary reference provided by the NIST.

2.1 Visual Features

Since the major focus of TRECVid 2007 [1] was to encourage video processing instead of keyframe processing, we made an attempt to track similar points from frame to frame to aid video retrieval. Before tracking similar points frame to frame, it was first necessary to find out similar points between a query image and a shot frame to start with. For this purpose we made use of the SIFT locations and SIFT features (Lowe, 1999, 2001, 2004). However, due to time constraints, we had to stick onto the keyframes, although we considered five keyframes in each shot irrespective of their lengths.

2.2 Textual Features

The videos were annotated manually with some additional labels used with that of the annotation labels provided for Trecvid 2006 collection.

3. Automatic Runs and Experimental Setup

We have submitted two fully automatic runs, which are described in the following subsections.

As mentioned in section 2.1, the major feature for both runs are the SIFT features. Although, TRECVid 2007 aimed at upbringing video processing research than mere still image (keyframe) processing, we still use the keyframes as the basic unit of processing. For UG_F_Sys1, we used one representative keyframe from the shots. On the other hand, for UG_F_Sys2, from each shot, irrespective of their length, five representative keyframes, at regular intervals depending on the shot length were used for retrieval.

Inorder to start a match, it was necessary to have a query, but since the query examples provided for each topic is more than one, selecting one query to start with was a question.

There were two ways we could think of resolving this. First being, to consider every example image as a query and then to select certain number of top ranked shots for each query as the final list of similar videos. The other possibility was, to compute a single query for each topic, from the given many example images for each topic. A very crude thought was to just add the example images and obtain one query representative. But, a study made on TRECVid 2006 collection, proved that just adding the images does not help. An alternative way of obtaining a single query, similar to that of a Hybrid image (Oliva et al., 2006) was also tried and observed to be a failure. Therefore, we reverted back to treating each example image as the query.

3.1 UG_F_Sys1

UG_F_Sys1 was based only on SIFT features. Only the example images given by NIST were used as query for each topic. For a few topics, for which an example image was not available, we randomly extracted keyframes from the example videos and used them as the query.

SIFT features from each query and image collection was extracted and matched. A dissimilarity value was computed using (1)

$$Dissimilarity(I_i, I_j) = 100 - \frac{\frac{(100*M)}{N_i} + \frac{(100*M)}{N_j}}{2} \quad .....(1)$$

Where,

M is the number of SIFT points matched between images I_i and I_j

N_i is the number of SIFT points in Image I_i and

N_j is the number of SIFT points in image I_j

3.2 UG_F_Sys2

For UG_F_Sys2, along with the setup used for UG_F_Sys1, we used manual annotation keywords. The keywords set contained the ones given for Trecvid2006 collection and some additional words. On specification of a query, a first level matching was done on a simple text search and all those videos having the query keywords were selected as the possible candidates and subjected to further processing. The dissimilarity formulation was slightly modified into a similarity measure as given in (2)

$$Similarity(I_i, I_j) = \frac{\frac{(100*M)}{N_i} + \frac{(100*M)}{N_j}}{2} \quad .....(2)$$

The SIFT features were matched with the keyframes K_i $1 \leq i \leq 5$, of the shots in sequence. Once, a considerably high match between the query and a keyframe K_i, is obtained, the matched SIFT points in K_i^{th} frame is tracked over the remaining keyframes k_j $i+1 \leq j \leq 5$, and a weight is assigned to the shot if more keyframes have high degree of match initially with the query and then with the preceding keyframes of the shot. The

similarity value is computed using (2), for the subsequent images.

Since we are using five keyframes from each shot far matching, inorder to select the topmost videos, the similarity values of all the five keyframes are summed up and a weight is added depending on the number of keyframes having a match with the query.

In addition to the above, the query set size was also increased. One keyframe was extracted for each example video given for the topic and were treated as the query along with the example images given by NIST.

Thus, along with the 73 example images provided by NIST, we extracted 132 keyframes from the example videos. The middle frame of the example video shot was selected as the keyframe automatically.

4. Results

Table 1: Overall experiment results

Run ID	MAP	P(10)	P(NR)	Recall
UG_F_Sys1	0.001	0.025	0.008	0.041
UG_F_Sys2	0.017	0.046	0.040	0.139

Table 2: MAP per topic

Topic	UG_F_Sys1	UG_F_Sys2
197	0.0005	0.0002
198	0.0005	0.0020
199	0.0022	0.0103
200	0.0016	0.0019
201	0.0003	0.0036
202	0.0002	0.0002
203	0.0001	0.0004
204	0.0002	0.0027
205	0.0004	0.0002
206	0.0040	0.0231
207	0.0018	0.0068
208	0.0000	0.0004
209	0.0014	0.0003
210	0.0001	0.0009
211	0.0002	0.0002
212	0.0001	0.0053
213	0.0008	0.0004
214	0.0000	0.0184
215	0.000 6	0.0016
216	0.0026	0.0031
217	0.0011	0.0126
218	0.0008	0.2639
219	0.0000	0.0015
220	0.0006	0.0547
All	0.0008	0.0173

The results of the submitted runs are given in Table 1, which compares mean average precision (MAP), precision at 10 (P(10)), precision at total relevant shots (P(NR)) and recall averaged over all topics. The MAP results per topic are shown in Table 2.

The run UG_F_Sys1, which was solely based on SIFT features was a disaster without any surprise.

However, UG_F_Sys2 which considered an added information of keywords/annotations, performed relatively better for a few topics, 220 (gray scale shots of a street with one or more buildings and one or more people), 206 (Shots with hills or mountains visible), 214 (shots of very large crowd of people filling more than half of field of view), and was at its best for 218 (people playing musical instruments).

For all other queries, the performance was bad and was exactly at the median line.

5. Conclusions

The Glasgow University team submitted two fully automatic runs. One of these runs was based only on SIFT features and the other based on a combination of SIFT and textual features, specifically, annotation/keywords. As expected, the run based solely on SIFT features performed poorly. However, the combination of textual features improved the results to some extent. We are currently analysing the results obtained and aim for a better retrieval system.

6. Acknowledgements

The research leading to this paper was supported by European Commission under contracts FP6-027026(K-Space) and FP6-027122(Salero).

The first author acknowledges the kind help rendered by Frank Hopfgartner at various stages of submission.

7. References

[1] Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVid. In proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330. DOI= http://doi.acm.org/10.1145/1178677.1178722.

[2] Lowe D. G., 2004, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision.

[3] Lowe D. G., 1999, Object recognition from local scale invariant feature, Proceedings of International Conference on Computer Vision.

[4] Lowe D. G., 2001, Local feature view clustering for 3D object recognition, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.

[5] Oliva A., Torralba A., Schyns P. G., Hybrid Images, 2006, ACM, 0730-0301/06/0700-0527.