

October 28, 2013

This report is in responding to the letter of doubt about BUPT INS 2012 results on August 18, 2013. Our preliminary check, made as soon as we received the letter, also showed that there existed some problems in those results. To be serious about TRECVID evaluation as well as about academic works, we decided to carry out an investigation on all the results we had submitted to TRECVID since our first participation in 2007. Now we report the investigation results as follows.

1. Scope

The investigation covers SBD 2007; HLF and CCD 2008; HLF, automatic Search, CCD and SED 2009; CCD, SIN, INS, KIS and SED 2010; INS, KIS and SED 2011; INS, KIS and SED 2012, 18 task • year in total .

2. Methods

The methods used in the investigation include:

- 1) Meeting with the original performers face to face or through Skype to review the procedure and the results of a particular task. This was done for every task • year, which involved 14 people.
- 2) Cross-checking the procedure and the results by people other than the original performers. This was done for some task • years, such as Search 2009 and INS 2010.
- 3) Re-performing the experiments of some task • years to check the results, such as on KIS 2011 and INS 2012.

3. Results

After careful examination and cross-checking, we assure that there is no problem in all the tasks except in INS 2012. The details on no-problem and having-problem tasks are respectively described in this section and Section 4.

1) Shot boundary detection (SBD 2007)

In 2007, we participated in TRECVID evaluation for the first time. We

developed a real-time video context and SVM-based system for robust and fast SBD. The CUT shot, GT shot and motion detectors were trained for making SBD decisions. Our 2007 notebook paper described it in detail.

This system showed good balance between accuracy and operational speed in TRECVID evaluation. Thus the algorithm developed in this task was first packaged into a commercial software for video summarization, and then became an essential tool applied to most of video processing tasks in our lab over the years.

2) **Content-based Copy Detection** (CCD 2008, 2009, 2010 and 2011)

We continuously attended CCD from the start of this task in 2008 to its ending in 2011. Different CCD systems were proposed each year. In 2008, only global visual features were extracted and applied to detect copy videos. In 2009, 2010 and 2011, we adopted similar detection frameworks, that is, SIFT with BoW-based method and a hierarchical matching scheme (points, segmentations, videos). In addition, audio features were added in 2010 and 2011.

The 2008 system achieved high precision and mean F1 scores, but relatively low recall especially for “video-in-video” transformation. The mean F1 and NDCR of 2010 and 2011 systems were better than the average of all participants in TRECVID evaluation. On the basis of these experiences, we developed a CCD software in a university-industrial cooperative project to retrieve commercials in TV streams. The successful application of this software supports the veracity of our submissions in CCD task.

3) **High-level feature extraction/semantic indexing** (HLF 2008 and 2009, SIN 2010 and 2011)

In 2008 ~ 2011, we proposed similar frameworks to detect concepts for shot key frames. Firstly, different local key points, i.e., SIFT, SURF in 2008 and 2009, CSIFT in 2010, and SIFT, CSIFT, Opponent-SIFT and RGB-SIFT in 2011, were extracted and projected to pre-trained codebooks by using BoW method. Various global color and texture features were also extracted. And then multi-SVM classifiers were trained as the concept detectors. Finally the detection results were fused by different later-fusion schemes to generate the final result.

The performances of these systems are all around the average of all participants in TRECVID evaluation.

4) **Surveillance event detection** (SED 2009, 2010, 2011 and 2012)

In this task, each year we focused on 4~6 events among PersonRuns, Pointing, ElevatorNoEntry, Take Picture, Opposing Flow, PeopleMeet, PeopleSplitUp, ObjectPut, Embrace, Pointing, and etc. Each event was detected by a SVM, a random forest or a cascade SVM-HMM classifier with different features such as HOG, MEH cubes, gray image cubes, MoSIFT and trajectory analysis. In SED 2012, we also submitted an interactive run.

The performances of PeopleSplitUp of 2011, and automatic and interactive runs of 2012, were around or above the average of all participants at that year, but other event detection results were not good.

5) **Automatic search 2009**

Our algorithm in automatic search 2009 composes concept-based retrieval, text-based retrieval and fusion of the two. The text-based query mainly relies on Lucene and WordNet, while the concept lexicon includes 64 concept detectors released by MediaMill, and 48 concept detectors developed by us at TRECVID 2008 for HLF task. We submitted 10 runs for this task.

The best run, F_A_N_BUPT-MCPR4, used all the above 112 concepts, and obtained the best MAP 0.131 in TRECVID evaluation. Careful cross-checking, as described in Section 2, was carried out in the investigation of this task. We found that the result obtained in F_A_N_BUPT-MCPR4 benefits from that (1) MediaMill concept detectors have good performance, and (2) some search topics in 2009 are the same as the concepts (high level features) in 2008 so our concept detectors were effectively utilized.

6) **Known-item search (KIS 2010, 2011 and 2012)**

In KIS 2010, a text-based search method was adopted. It mainly included four parts: 1) OCR for video clips; 2) text pre-processing, which contained spelling check and correction, lemmatization by Porter stemmer, and keyword extraction; 3) text-based retrieval according to Lucene indexing; 4) re-ranking based on visual contents such as face, color, and black and white.

In KIS 2011 and 2012, expect a text-based search method which was the modified versions of 2010, we also proposed a novel method based on visual attention model. In this method, a query topic was first parsed by a text analyzer to produce several image cues, and then the cue-based bottom-up saliency map and the top-down cue-guided concept/object detection were fused and refined with the aid of context

cues. Although MAPs of this method were quite low in TRECVID evaluation, the performance improvement year after year may show its potential of being an alternative approach for text-image-based retrieval.

The performances of our text-based search were good in these three years, and especially the best run, F_A_YES_MCPRBUPT1_1 of 2011, achieved the best MAP in TRECVID evaluation. To verify these results, we re-performed the experiments for KIS 2011 task in the investigation, and the results showed that our system can achieve the performance reported in F_A_YES_MCPRBUPT1_1. The major improvements of the later systems over the 2010 one, which offered the good retrieval results, were that (1) new rules and strategies for key word selection were proposed in text pre-processing; (2) query scores for various keyword collocation with different regular expressions were ranked and the best one was chosen as the results in retrieval module; (3) a part of the software in our attention-model-based method, which contained 32 image cues, was adopted for re-ranking.

7) Automatic instance search (INS 2010 and 2011)

In 2010 and 2011, we adopted similar automatic INS systems consisting of visual query pre-processing, feature extraction, key frame retrieval including face-based, body color-based and global image-based retrieval modules, result fusion and re-ranking.

The performances of the systems in these two years were relatively good in TRECVID evaluation. In the investigation, we carried out a cross-checking on INS 2010, one of the two task • years, to exam its results, and found no problem in them. The contributions to the relatively good results mainly came from that (1) different features were used in different retrieval modules, and different retrieval modules were invoked for different search topics; (2) body color-based retrieval, according to ROIs (region of interest) defined via face detection, was effective for some search topics.

4. Problem

There are problems in automatic instance search (INS) 2012. For this task, the investigation was also started with meeting the original performers to review the procedure and the results of that year. During the meeting, one student (original performer) admitted that under the pressure of not being able to give performance as good as achieved by our lab in past years, he illegally used an interactive algorithm, which was developed in the lab for another project, on about half of the queries in

automatic search without professors' consents.

In order to verify what he said and to see actual performance of the automatic search algorithm, we repeated the experiments. The experiments started from extracting key frames and features because 2012 data had not been kept. Due to the same reason, the dictionary used this time was a 60K one generated from several public datasets and commonly used for many projects in our lab, while the dictionary used in our INS 2012 was a 50K one generated from TRECVID dataset.

Our algorithm in INS 2012 contained two kinds of features, global one and regional one. Both features were generated with SIFT, soft assignment and BoW. When calculating the similarity we gave more weight to the regional feature. RANSAC was used to do the result re-ranking. In addition, two special treatments were employed in the algorithm. One was for logo query, and another for people query. For logo query, we added on a detection-based method. A cascaded classifier for a logo was trained based on logo images from publicly available datasets, e.g. MICC_logo (Coca-cola, Pepsi, MacDonald) and Belga-Logo (Mercedes and Puma) datasets. Then the classifier was used to locate possible logo regions on the reference image and the score was computed by comparing the detected candidate region with a chosen logo template. Finally, we combined the detection score together with the retrieval score described above to give the final result. For people query, a face detection algorithm was used to filter out the retrieved images without faces.

In the re-examination experiments, 10 (No.9053, 9055, 9056, 9057, 9061, 9063, 9064, 9065, 9067, 9068) of total 21 search topics can obtain the results of our INS 2012 submission with automatic algorithm, while the rest of them need the help of interactive algorithm to achieve the results of our INS 2012 submission. This matches what the student has admitted.

Two suspected topics, No.9067 and No.9048, are further explained as follows.

The doubt on 9067 is that the background of the image on the first place of the result list (see below) looks different from the topic images (dark background).

Actually, the logo detection took effect here. A MacDonald logo classifier was trained with 52 positive examples and 100 negative examples from MICC-logo dataset. In the re-examination experiments, the following image was retrieved and ranked on the 1st place with automatic algorithm, the same situation as in our INS 2012 submission.



The doubt on 9048 is that the first 8 results contain "Mercedes star". However, the 2nd (FL000037674) and 5th (FL000024758) results are not regarded as correct results by NIST, since the 2nd video is not clear and the "Mercedes star" only appear in one frame of 5th video.

In the re-examination experiments, 4 (including FL000037674) of the 8 were ranked in top 20, and other 2 in top 500, with automatic algorithm. To achieve the results of our INS 2012 submission, interactive algorithm had to be applied. However, this time we did not find FL000024758 even with interactive algorithm. This may be due to the fact that key frames selected in the two experiments could be different and only few frames contain "Mercedes star" in FL000024758.

5. Conclusions

Rigorous scholarship has been being our tenet since the lab was established in 1989. We hate any kind of misconducts. Unfortunately, it happened to our lab this time. We apologize for the negative influence to TRECVID evaluation brought by our mistakes. Learning from this, we have tightened the lab's disciplines in order to guarantee that no similar things happen again. We will, as always, participate in TRECVID evaluation, and wish we could be together with academic colleagues to promote technologies in the area of content-based analysis through TRECVID evaluation platform.