Telecom Italia at TRECVID2014 - Instance Search Task*

Luca Bertinetto[1], Massimo Balestri[1], Skjalg Lepsøy[1], Gianluca Francini[1], Enrico Magli[1], and Miroslaw Bober[2]

[1]Telecom Italia, Joint Open Lab VISIBLE
[2]Centre for Vision, Speech and Signal Processing, University of Surrey

Abstract

MPEG CDVS is a forthcoming standard for representing images for visual search, both as queries and as database entries, with very short files, on the order of a kilobyte. The key advantages are interoperability and compactness. Interoperability ensures that data and equipment from different sources are mutually compatible, compactness allows for inexpensive object retrieval in terms of transmission and storage.

In the Instance Search Task of TRECVID2014, CDVS performs well on rigid and detailed objects. For queries that depict persons, animals, or objects of poor detail, CDVS may not be sufficient as a standalone technique.

1 Introduction

This document describes Telecom Italia's submission to TRECVID2014 (1) *Instance Search* task, in collaboration with the University of Surrey. The workhorse of the pipeline is MPEG-7 CDVS (Compact Descriptors for Visual Search, in its version Test Model 10), soon to become an ISO/IEC standard (end 2014/beginning 2015) (2).

CDVS specifies the extraction of parameters from single images, as well as the binary representation of these parameters. The software that embodies the standard also provides exemplary modules for image comparison, index construction, and object retrieval. In the experiment for TRECVID, the two latter modules are used.

*This work was supported by the European Community in the context of the BRIDGET (FP7-ICT- 610691) project. Further information can be found at http://ict-bridget.eu/.

2 The CDVS standard

The ISO/IEC standard MPEG-7 CDVS[1] provides an image description tool designed for efficient and interoperable visual search applications. This tool allows for matching and search of visual content like objects, landmarks, and printed documents. It is robust to partial occlusions as well as changes in viewpoint, camera parameters, and lighting conditions.

The CDVS method describes a single image with a very short file. This descriptive file, called a *descriptor*, can be used both as a query and as a representation of an image in a searchable index. As the descriptors are short (sizes among 512B, 1KB, 2KB, up to 16KB), the queries are simple to transmit and store. If the indexed image collection is not too large, the index may be kept in the memory of a single smartphone. For example, an index for 1000 images typically occupies 2MB, a memory footprint that compares well to the median footprint of 4.8 MB found in a survey of visual search applications (indices mostly excluded) by Chandrasekhar *et al.* (3).

A CDVS descriptor contains three types of information: the coordinates of interest points, local features for the same interest points, and a global descriptor (2). The interest points are detected through an analysis of a scale space representation of an image and a subset selection process. The local features are derived from histograms of gradients around each interest point, and the global descriptor is a scalable Fisher vector for a mixture model of the set of local features. The global descriptor is designed for fast search in large image collections: a 'global score' is assigned to each entry in the collection, and these scores are sorted such that the top entries form a shortlist of candidates. The descriptors in the shortlist are then compared to the query descriptor by matching local features and checking the geometric consistency of matched interest points, producing 'local scores'. The local scores for the shortlist are sorted to provide the final ranked output images.

This technique is suited for retrieval of rigid objects with some visible surface pattern. It recognizes individual instances and not categories of objects.

2.1 Training data

There are five trained subsystems in the CDVS tools of descriptor extraction, matching and retrieval. These subsystems regard: local feature selection, local feature compression, the global descriptor, location coding context, and weighted matching of the local features.

However, the CDVS tools are trained once and for all. The training data has no overlap with the video material in the instance search task of TRECVID.

The CDVS specification does however allow for tuning some parameters, such as decision thresholds and number of local features contained in a descriptor. None of these were changed with respect to the setup used for experiments in MPEG.

[1] Its long name is: Information technology - Multimedia content description interface - Part 13: Compact descriptors for visual search, and its alphanumeric code is ISO/IEC 15938/13.

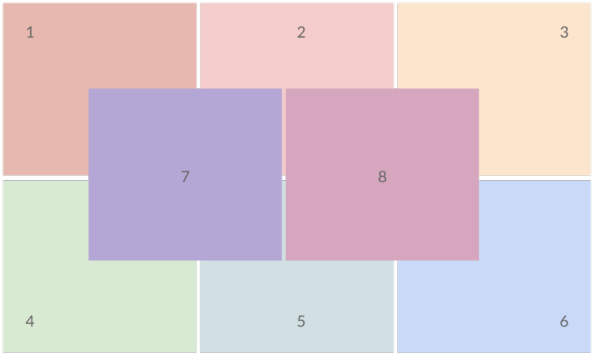

Figure 1: Each keyframe is cropped in 8 sub-frames, that are used to augment the dataset in order to be more efficient in retrieving small objects.

3 Pipeline

3.1 Keyframe extraction and database creation

In order to generate the database, we extracted approximately 1.2 million keyframes from the 464 hours of video of *BBC EastEnders*. Instead of using a constant sampling (e.g. one keyframe per second), we decided to sample with an higher frequency in the portion of video with more movement. The reason behind this choice is that during a fast moving scene several potential targets can quickly appear and disappear, and thus a more thorough representation should be enforced. We simply compare the sum of absolute differences of the color histograms of two consecutive frames with a threshold value, obtaining a *micro-shot* segmentation within each shot [2]. Then, we extract the median frame of each *micro-shot* as keyframe. Using TRECVID2013 ground truth and keeping the number of extracted keyframes constant, we verified that this representation technique achieves better results than sampling at a constant rate.

Eventually, we augment the dataset cropping 8 sub-frames from each keyframes as showed in Figure 1. This helps to retrieve better small object that occupy only a small portion of the frame.

3.2 Late fusion

For each topic we query the system twice with all the four different images provided: one with the full image and one only considering the area delimited by the bounding polygon. Sometimes indeed, the *context* of the object query is much more informative than the object itself, thus the surrounding space could bring important and complementary information. In Figure 3, for example, the local features belonging only to the query objects (*i.e.* the two obelisks) are not very informative. In this case the query object is somehow related to the furniture present in the room,

[2] We used the open source implementation of `github.com/johmathe/Shotdetect`

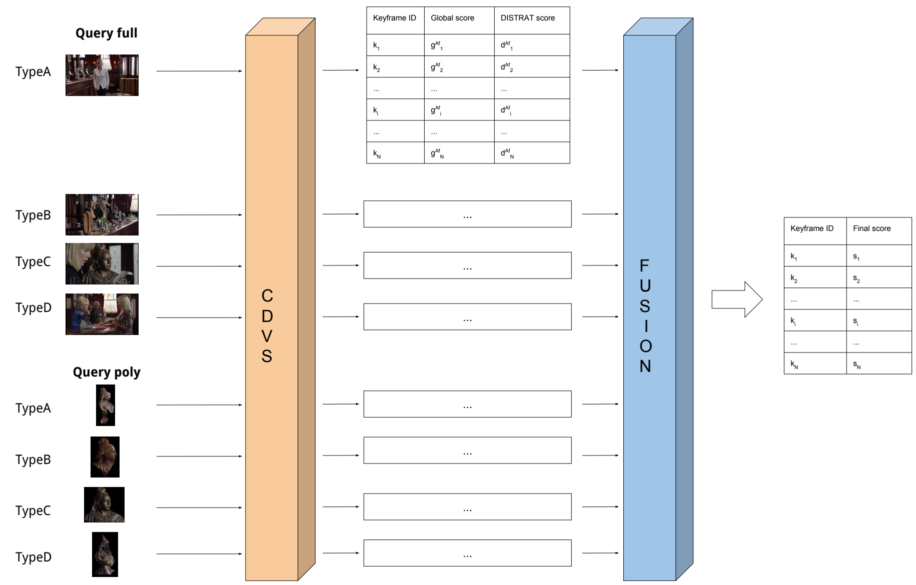

Figure 2: High level outline of the general pipeline adopted for TRECVID2014 Instance Search task. For each of the 30 given topic, 8 images are used as queries to the CDVS-TM and then the 8 partial results are combined to obtain the final scores.

so that if the obelisk appears, it is quite likely that also the painting with the flowers appear. On the other hand, there are also cases in which the query object is completely uncorrelated with the surrounding space, hence it is important to use only local features belonging to the very object in order not to mislead the system.

We observed that using both results of "bounded" and "full" query image it is possible to obtain complementary results that, once conveniently combined, entail a higher performance.

After having queried the database eight times for every given topic, for each of the eight images we obtain a list of retrieved results like the one illustrated in the table of Figure 2.

First, we normalize all the scores in the interval $[0, 1]$. Then, for each retrieved keyframe k_i^q from query image $q \in [1, 8]$ (4 for "full" and 4 for "bounded" queries) we combine the *global score* g_i^q and a *local score* d_i^q in a score s_i^q. Given G^q the number of nonzero global scores and D^q the number of nonzero scores for image query q, we experiment with three different *data-fusion* techniques.

Run 1 and 4. For each keyframe, we consider only its top-retrieved subframe, discarding all the others. To obtain the final score, we apply

$$s_i^q = d_i^q \sqrt{G^q} + g_i^q \sqrt{D}^q,$$

that manages to compensate the high number of zero-valued local

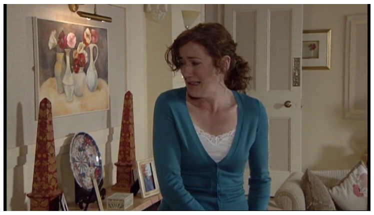

Figure 3: Example image in which the context of the query object (*i.e.* the obelisk) significantly contributes to the effectiveness of retrieval. Programme material copyrighted by BBC.

scores.

Run 2 and 5. Again, we consider only the top-retrieved sub-frame. We simply use CombSUM (4) to combine global and local scores and obtain S_i:

$$S_i = g_i^q + d_i^q.$$

Run 3 and 6. For each keyframe, we obtain the partial scores g_i^q and d_i^q by summing all the scores obtained by its sub-frames. Then, global and local scores are combined with CombSUM.

Eventually, for each keyframe i, the final score S_i is given by simply combining the S_i achieved by the 8 query images

$$S_i = \sum_{q=1}^{8} s_i^q.$$

3.3 Results and discussion

3.3.1 Performance

Table 1: mAP for all our six runs.

Run #	**Examples set**	**mAP**
1	D	0.138
2	D	0.136
3	D	0.126
4	C	0.126
5	C	0.124
6	C	0.118

Table 1 displays the performance in mean average precision of all our six runs, while the plot of Figure 4 shows the per-topic performance of *Run 1* in comparison with the other teams.

Our purpose was to test CDVS-TM *as-it-is* in the broadly scoped TRECVID *instance search* task. As expected, it performs well with queries that depict rigid and detailed objects. In particular, in topics 9099, 9111, 9108 and 9118 our contribution scored among the best.

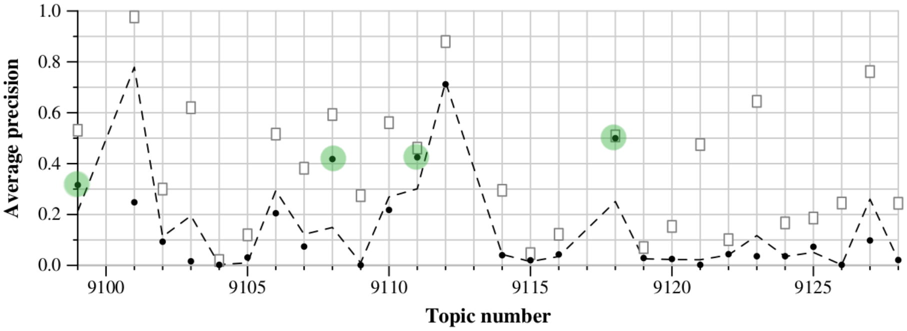

Figure 4: Run #1 score (dot) versus median (dashed line) versus best (box) by topic. The green circles indicate the topics for which our system performed particularly well. The query images for these topics depict rigid objects with detailed surface patterns.

On the other hand, our system performs poorly for the queries that portray people (9104, 9115, 9116, 9119, 9124). Moreover, many topics depicted objects with almost no details, like the ketchup container (9103), the mailbox (9114), the wooden bench (9120), the red vest (9121) or the plastic kettle (9123). Such objects are hard to recognize with a system based on SIFT-like local features like CDVS.

A comment is due for the washing machine (9101). For this topic our system retrieved a very large number of images depicting washing machines. We speculate that many of these washing machine images were not contained in the ground truth, perhaps for the reason that one specific machine was to be detected and not the other ones, albeit quite similar.

3.3.2 Timing

Our "D" runs required an average of 486 seconds per topic. Considering that each topic comprises eight queries to the system (see Section 3.2), the time required to reply to a single query image was about 1 minute. In case of need, the time can be widely reduced avoiding database redundancy. We experimented that, without using subframes, the per-query-time is reduced to just 3 second, with a drop of about 15% in mAP performance.

4 Conclusion

Our retrieval system based on CDVS Test Model performed well on those topics for which it is designed, namely rigid objects that have a certain amount of surface detail. Other types of objects were found to be overly challenging. Our current pipeline would enormously benefit from adding other types of local features and face recognition capabilities. Eventually, the partial results can be easily merged using late fusion techniques.

References

[1] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quéenot, “Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics,” in *Proceedings of TRECVID 2014*, NIST, USA, 2014.

[2] “Study text of ISO/IEC DIS 15938-13 Compact Descriptors for Visual Search.” ISO/IEC JTC1/SC29 (MPEG), 2014. Output document W14681.

[3] V. Chandrasekhar, D. Chen, G. Takacs, S. Tsai, M. Makar, R. Grzeszczuk, and B. Girod, “Sizes of typical state-of-the-art visual search applications.” ISO/IEC JTC1/SC29 (MPEG), 2012. Input document M23581.

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