2011

Bag-of-visual-words vs global image descriptors on two-stage multimodal retrieval

Zagoris, Konstantinos

ACM SIGIR 2011

Bag-of-Visual-Words vs Global Image Descriptors on Two-Stage Multimodal Retrieval

Konstantinos Zagoris Savvas A. Chatzichristofis Avi Arampatzis
Department of Electrical and Computer Engineering
Democritus University of Thrace, Xanthi 67100, Greece
{kzagoris,schatzic,avi}@ee.duth.gr

ABSTRACT

The Bag-Of-Visual-Words (BOVW) paradigm is fast becoming a popular image representation for Content-Based Image Retrieval (CBIR), mainly because of its better retrieval effectiveness over global feature representations on collections with images being near-duplicate to queries. In this experimental study we demonstrate that this advantage of BOVW is diminished when visual diversity is enhanced by using a secondary modality, such as text, to pre-filter images. The TOP-SURF descriptor is evaluated against Compact Composite Descriptors on a two-stage image retrieval setup, which first uses a text modality to rank the collection and then perform CBIR only on the top-K items.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*retrieval models, search process*

General Terms: Measurement, Experimentation, Theory

Keywords: Image Retrieval, Bag-Of-Visual-Words, Two-Stage Retrieval

1. INTRODUCTION

In Content-Based Image Retrieval (CBIR) images are represented by either global or local features. Global features are capable of generalizing an entire image with a single vector, describing color, texture, or shape. Local features are computed at multiple points of interest on an image and are capable of recognizing objects.

Compact Composite Descriptors (CCDs) [4] are global image features capturing more than one type of information at the same time in a very compact representation. Their retrieval quality has so far been evaluated in several benchmarking databases and is found to be better than other descriptors such as the MPEG-7 descriptors.

SURF local features are among the best interest-point descriptors currently available. They have been shown to outperform other well-known methods based on interest points such as SIFT and GLOH. Nevertheless, in large-scale CBIR, it is clear that using the SURF descriptor is storage-wise infeasible [6].

Bag-Of-Visual-Words (BOVW) [5] is a representation of images which is built using a large set of local features. They are inspired by the bag-of-words models in text retrieval, where a document is represented by a set of distinct keywords. Analogously, in BOVW models, an image is represented by a set of distinct visual words derived from local features. The most modern implementation of BOVW suitable for a wide range of CBIR applications is the TOP-SURF [6] descriptor. TOP-SURF combines interest points with visual words, resulting in a high performance compact descriptor.

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

The TOP-SURF descriptor initially extracts SURF local features from images and then groups them into a desired number of clusters. Each cluster can be seen as a visual word. All visual words are stored in a visual dictionary. Next, tf.idf weighting is applied in order to assign a score to all the visual words. The TOP-SURF image descriptor is created by choosing the top-scoring visual words.

Nowadays, information collections are multimodal and large (for example Wikipedia), where a single topic may be covered in several languages and include non-textual media such as image, audio, and video. In an image retrieval system where users are assumed to target visual similarity, all modalities beyond image can be considered as secondary; nevertheless, they can still provide useful information for improving CBIR.

In [2], we experimented with two-stage image retrieval from a large multimodal database, by first using a text modality to rank the collection and then perform CBIR only on the top-K items. The approach employed CCDs and was found to be significantly more effective than the text-only and image-only baselines when K was dynamically calculated with respect to the underlying query generality. Traditionally, the method that has been followed in order to deal with multimodal databases is to search the modalities separately and fuse their results. While fusion has been proven robust, we also found that two-stage is more effective than fusion [3]. Furthermore, a two-stage approach has an efficiency benefit: it cuts down greatly on expensive image operations.

The BOVW paradigm is fast becoming a widely used representation for CBIR, mainly because of its better retrieval effectiveness over global feature representations on collections with images being near-duplicate to queries. In this experimental study, we evaluate the performance of the BOVW approach in comparison to CCDs in a multistage multimodal setup. We intend to check whether the aforementioned reason holds also in such setups, where CBIR is performed on a pre-filtered set of images with a high probability of relevance. Given that the high relevance of such a subset is based on the text modality, it is likely to consist of more diverse images than the top-retrieved images of CBIR systems.

2. TOP-SURF VS CCDS

We experimented with the ImageCLEF 2010 Wikipedia test collection, which consists of 237,434 images associated with noisy and incomplete user-supplied textual annotations. There are 70 test topics, each one consisting of a textual and a visual part with one or more example images. The topics were assessed by visual similarity to the image examples. The collection is one of the largest benchmark image databases for today's standards. It is also highly heterogeneous, containing color natural images, graphics, grayscale images, etc., in a variety of sizes.

We indexed the images with two CCDs: the Joint Composite De-

scriptor (JCD) and the Spatial Color Distribution (SpCD) descriptor [4]. We also indexed the images with the TOP-SURF descriptor, employing two visual-word dictionaries: one with 10,000 and the other with 200,000 visual words. For CBIR we used the JCD, SpCD, and TOP-SURF, separately, as well as a late fusion setup of JCD and SpCD explained next.

Let i be the index running over example images ($i = 1, 2, \ldots$) and j running over the visual descriptors ($j \in \{1, 2\}$). Thus, DESC_{ji} is the score of a collection item against the ith example image for the jth descriptor. We normalize DESC_{ji} values with MinMax, taking the maximum score seen across example images per descriptor. Assuming that the descriptors capture orthogonal information, we add their scores per example image. Then, to take into account all example images, the natural combination is to assign to each collection image the maximum similarity seen from its comparisons to all example images; this can be interpreted as looking for images similar to *any* of the example images. Summarizing, the score s for a collection image against the topic, for the JCD/SpCD fused setup, is defined as:

$$s = \max_i \left(\sum_j \mathrm{MinMax}(\mathrm{DESC}_{ji}) \right) \quad (1)$$

With the same reasoning, the $\max_i$ is applied also in the TOP-SURF runs to handle multi-image topics.

For text indexing and retrieval, we employed the Lemur Toolkit V4.11 and Indri V2.11 with the tf.idf retrieval model. We used the default settings that come with these versions of the system except that we enabled Krovetz stemming. We indexed only the English annotations, and used only the English query of the topics.

First, the collection was ranked with the secondary text modality, and then the top-K results were re-ranked by the primary visual modality using CBIR methods based on the aforementioned descriptors. The threshold K was calculated dynamically per query using the Score-Distributional Threshold Optimization (SDTO) [1]. The method normalizes retrieval scores to probabilities of relevance (prels), enabling the the optimization of K by thresholding on prel. We report results for three prel thresholds, i.e. 0.800, 0.500, and 0.333; these were the best three performers in [2].

We evaluated on the top-1000 results with MAP, precision at 10 and 20. We tested the results for statistical significance against the text-only baseline; image retrieval based on the text queries and annotations was found to perform much better, with a wide margin, than CBIR-only in the same setup [2]. For measuring efficiency, we report the average matching time per topic. The results are presented in Table 1.

For all θ, the CCDs perform similarly (JCD) or significantly better (SpCD and JCD/SpCD) than the text-only baseline, while the TOP-SURF descriptor shows significant drops in effectiveness irrespective of dictionary size. The differences in effectiveness of CCDs and TOP-SURF are larger in early precision than in MAP. We also observe that the TOP-SURF effectiveness degrades with increased dictionary size. Furthermore, TOP-SURF is more sensitive to the choice of θ: as θ decreases (i.e. for larger Ks), effectiveness deteriorates faster than this of the CCDs.

Efficiency-wise, the experimental results show that although the TOP-SURF uses a speedy matching algorithm, it still cannot match the speed of the global descriptors.

descriptor	θ	MAP	P@10	P@20	avg. time (sec.)
text-only	-	.1293	.3614	.3307	-
TOP-SURF (10,000)	.8000	.1156$^{\triangledown}$	.3743-	.3264-	.4359
	.5000	.0999$^{\blacktriangledown}$	.3200-	.2850$^{\triangledown}$	.7748
	.3333	.0893$^{\blacktriangledown}$	.2686$^{\triangledown}$	.2557$^{\triangledown}$	.7793
TOP-SURF (200,000)	.8000	.1149$^{\triangledown}$	.3300$^{\triangledown}$	.2979$^{\triangledown}$	.7714
	.5000	.1030$^{\blacktriangledown}$	.2857$^{\triangledown}$	.2579$^{\blacktriangledown}$	.6665
	.3333	.0956$^{\blacktriangledown}$	.2629$^{\triangledown}$	.2229$^{\blacktriangledown}$	.4185
JCD/SpCD	.8000	**.1428**$^{\blacktriangle}$	**.4443**$^{\blacktriangle}$	**.3857**$^{\blacktriangle}$	.0360
	.5000	.1405$^{\triangle}$	.4357$^{\blacktriangle}$	.3821$^{\triangle}$	.0432
	.3333	.1403$^{\triangle}$	.4357$^{\triangle}$	.3807$^{\triangle}$	.3632
JCD	.8000	.1348$^{\triangle}$	.4271$^{\triangle}$	.3743$^{\triangle}$	**.0095**
	.5000	.1305^{-}	.4171$^{\triangle}$	.3693$^{\triangle}$	.0110
	.3333	.1315^{-}	.4114^{-}	.3707^{-}	.2870
SpCD	.8000	.1342$^{\triangle}$	**.4443**$^{\blacktriangle}$	.3771$^{\triangle}$	.0363
	.5000	.1302$^{\triangle}$	.4329$^{\triangle}$	.3693$^{\triangle}$	.0411
	.3333	.1307$^{\triangle}$	.4286$^{\triangle}$	.3743$^{\triangle}$	.0457

Table 1: Retrieval effectiveness and matching time. The best results per measure and retrieval type are in boldface. Significance-tested with a bootstrap test, one-tailed, at significance levels 0.05 ($^{\triangle\triangledown}$), 0.01 ($^{\blacktriangle\blacktriangledown}$), and 0.001 ($^{\blacktriangle\blacktriangledown}$), against the text-only baseline. Experiments contacted on a Pentium Dual-Core E2200 (2.4 GHz) with 4GB memory.

3. CONCLUSIONS

We investigated the performance of BOVW models, specifically the TOP-SURF image descriptor, in a two-stage multimodal retrieval setup, in comparison to CCDs. We found that CCDs are more effective, as well as faster in matching speed, than TOP-SURF. Although BOVW models are currently trendy because of their ability to recognize objects and retrieve near-duplicate (to the query) images, this advantage over global features such as CCDs is diminished when visual diversity is enhanced by using a secondary modality, such as text, to pre-filter images. In practice this means that, applications like Google Goggles, where a user is querying an image in order to recognize a logo or a famous painting, BOVW models should be more effective. But in applications like Google Similar Images, where images are pre-filtered by text similarity, global features should be more suitable.

4. REFERENCES

[1] A. Arampatzis, J. Kamps, and S. Robertson. Where to stop reading a ranked list? Threshold optimization using truncated score distributions. In *SIGIR*, pages 524–531. ACM, 2009.

[2] A. Arampatzis, K. Zagoris, and S. A. Chatzichristofis. Dynamic two-stage image retrieval from large multimodal databases. In *ECIR*, volume 6611 of *Lecture Notes in Computer Science*, pages 326–337. Springer, 2011.

[3] A. Arampatzis, K. Zagoris, and S. A. Chatzichristofis. Fusion vs. two-stage for multimodal retrieval. In *ECIR*, volume 6611 of *Lecture Notes in Computer Science*, pages 759–762. Springer, 2011.

[4] S. A. Chatzichristofis, A. Arampatzis, and Y. S. Boutalis. Investigating the behavior of compact composite descriptors in early fusion, late fusion, and distributed image retrieval. *Radioengineering*, 19 (4):725–733, 2010.

[5] O. G. Cula and K. J. Dana. Compact representation of bidirectional texture functions. In *CVPR (1)*, pages 1041–1047, 2001.

[6] B. Thomee, E. M. Bakker, and M. S. Lew. Top-surf: a visual words toolkit. In *ACM Multimedia*, pages 1473–1476, 2010.