Proceedings of the International Conference on Machine Vision and Machine Learning
Prague, Czech Republic, August 14-15, 2014
Paper No. 139

Improving Performances of BoW-based Image Retrieval by Using Contextual Keypoint Descriptors

Andrzej Śluzek
Khalifa University, Department of Electrical and Computer Engineering
P.O. Box 127788, Abu Dhabi, UAE
andrzej.sluzek@kustar.ac.ae

Ryszard Kozera
Warsaw University of Life Sciences - SGGW
ul. Nowoursynowska 159, 02-776 Warszawa, Poland
ryszard_kozera@sggw.pl

Abstract –The paper reports an improved method of content-based image retrieval using a well-known method of bag-of words (BoW). Words built over descriptors of popular affine-invariant keypoint detectors (Harris-Affine and Hessian-Affine are exemplary choices) are used. What is novel, however, is the number of descriptors (i.e. the number of words) representing individual keypoints. Instead of SIFT (or another alternative descriptors SIFT-like descriptors representing both visual properties of keypoints and their local configurations are proposed (adopted from our previous works). In average, each keypoint has 10-15 such descriptors, but the increased size of BoW representation is in our opinion acceptable because of significant performance improvements in BoW-based image retrieval, as shown in a feasibility study on a popular benchmark dataset. Such an improvement is possible because very large vocabularies can be built over the proposed descriptors without compromising the sensitivity of words to minor geometric and photometric distortions.

***Keywords*:** Bag-of-words, Image retrieval, Affine-invariant keypoints, Keypoint descriptor, SIFT.

1. Introduction

Keypoint-based image (and sub-image) retrieval is one of the standard techniques in CBVIR. Keypoint similarities (in practice defined by identical visual words) indicate similar image fragments, which in turn help to identify near-duplicate images or to identify/localize near-duplicate image areas (in sub-image retrieval). Because the size of processed datasets can be very large (in particular in visual browsing, e.g. Chum, Matas, 2010, Jegou et al., 2010 or Stewenius et al., 2012) scalability and computational complexity are the fundamental factors in the underlying algorithms.

One of the standard techniques is *bag-of-words* (e.g. Csurka et al., 2004) where the visual similarity is represented by the similarity of word distributions (i.e. sparse histograms) over the compared images. The BoW model ignores the spatial locations of keypoints (which is its major disadvantage) so that most works introduce the geometric/configurational verification step to identify actually similar (sub-)images within the candidates preliminarily found by BoW. This is a computation-intensive task and many attempts have been reported (e.g. Chum et al., 2009, Jegou et al., 2010, Tolias and Jegou, 2014) to simplify it, although the shear presence of such verification limits the size of databases and/or affects the time-efficiency of the retrieval process.

In this paper, we propose a modification of the BoW model, where the same general principles are applied. However, the novel elements are: (A) very large vocabularies of (contextual) visual words are used, (B) the numbers of words describing individual keypoints are larger (typically 10-15 words), and (C) the same mechanisms are used both for the BoW pre-retrieval and for the subsequent verification

(which does not require any geometric analysis). Points (A) and (B) are briefly presented in Section 2. Section 3 explains details of BoW and Point (C). Illustrative exemplary results are given in Section 4.

2. Contextual Keypoint Descriptors

2. 1. Keypoint Neighbourhoods

The most informative and visually prominent image fragments are typically combinations of region features and contour features. Hessian-Affine (*hesaff*) and Harris-Affine (*haraff*) keypoints are popular examples of such features. Thus, we propose to consider *hesaff* keypoints in conjunction with neighbouring *haraff* keypoints (a region feature with surrounding contour features) or another way around (a contour feature with surrounding region features). Since the affine-invariant keypoints are represented by ellipses, such conjunctions can be formalized as follows:

A neighbourhood of a central *heraff* (*hasaff*) keypoint K with the ellipse E_K consists of M context *haraff* (*hesaff*) keypoints L_i with E_i ellipses for which:

1. The Mahanalobis distances D_M between the keypoints are:

$$1/\sqrt{2} \le D_M(K, L_i) \le 2 \quad (1)$$

   where the unit distance is defined by the shape of E_K ellipse, i.e. only keypoints within a predefined distances from K are included..

2. The areas of E_0 and E_i ellipses satisfy:

$$0.5 area(E_K) \le area(E_i) \le 1.5 area(E_K) \quad (2)$$

i.e. only keypoints of relatively similar sizes can be included (for effective scale invariance). It can be noted that the ellipses in Fig. 1 satisfy both (1) and (2).

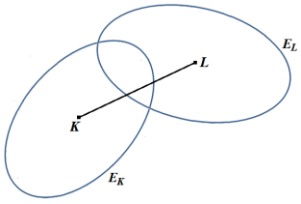

Fig. 1. Exemplary central keypoint K and one of its neighbours (context keypoints) L.

The average size of such neighbourhoods is 8-10 (both for *haraff* and *hesaff* keypoints); it is also possible to limit the maximum size (e.g. up to 20 keypoints).

2. 2. Keypoint Description

SIFT (see Lowe, 2004) is apparently one of the most popular keypoint descriptors. Therefore, the proposed method is also based on similar principles, but the actual descriptor is a concatenation of three SIFT-like vectors. Altogether, the proposed *contextual* SIFTs (CONSIFTs) are defined as follows (more details in Sluzek, 2014):

*Given a (central) keypoint K and a (context) keypoint L (with E_K and E_L ellipses, see Fig.*1*), a* 384*D (*3×128*D) CONSIFT descriptor of K in the context of L is obtained by concatenating (a) the original SIFT computed over E_K ellipse, (b) SIFT computed over E_K ellipse but using $(\overrightarrow{K,L})$ vector as the reference orientation and (c) SIFT computed over E_L ellipse with $(\overrightarrow{L,K})$ vector as the reference orientation.*

Thus, the number of CONSIFTs for an individual keypoint is determined both by the number of maxima in the gradient histogram over the keypoint ellipse (part (a)) and by the number of context

keypoints in its neighborhood. Since the number of standard SIFTs for a single keypoint is (according to our earlier studies) 1.4 in average, and the size of neighbourhood is typically 8-10 (see above), the numbers of CONSIFTs are usually within 10-15 range.

Moreover, CONSIFT descriptors can be easily quantized into words using the original SIFT vocabulary. If each of (a), (b) and (c) parts is represented by an *N*-word SIFT vocabulary, the CONSIFT vocabulary is simply a Cartesian product of N^3 size.

Thus, it is possible to build huge CONSIFT vocabularies using small SIFT vocabularies, For example, 1 billion CONSIFT words results from just 1000 SIFT words. Those huge vocabularies combine a high level of distinctiveness (because of their size) with insensitivity to minor distortions (because of a very coarse quantization of individual components).

3. CONSIFT words in BoW and Beyond

BoW model based on CONSIFT words differs from the SIFT-based approach mainly by the size of vocabulary. SIFT vocabularies reach at most a few million words (e.g. Nister and Stewenius, 2006) and larger sizes usually deteriorate performances, while 1 billion is a minimum practical size of the CONSIFT vocabulary. Therefore, CONSIFT histograms are extremely sparse (even though each keypoint is represented by several words) and in many cases the BoW similarity can be the arithmetic sum of AND operations over the histogram bins. Actually, we use a slightly different but also simple measure of histogram similarity, i.e.

$$s(H_1, H_2) = \sum_{x \in Voc} \min(x_1, x_2) \tag{3}$$

where x_1 and x_2 indicate the frequency of x word in both histograms, correspondingly.

The *proof-of-concept* experiment reported in Section 4 shows that CONSIFT-based BoW approach retrieves similar images more reliably the SIFT-based counterparts.

3. 1. Configurational Verification Using CONSIFT Words

The most typical methods verifying similar images retrieved by BoW apply geometric constraints (e.g. RANSAC, the Hough transforms, geometric hashing – Chum et al., 2009, Stevenius et al., 2012, etc.) over groups of preliminarily matched keypoints. In spite of the proposed improvements, these are always costly operations limiting the speed and/or the size of datasets. We propose to employ individual keypoint matches using multiple CONSIFT words. In other words, two keypoints are considered a match if they share at least *P* CONSIFT words in their description. This simple method provides (using the CONSIFT vocabulary of 1 billion words) a very high precision and a reasonably high recall (exceeding 80% and 40%, correspondingly, for $P = 2$) over the standard benchmark available in Web-1. No other work reports such performances, if not supported by the actual geometric/configurational verification.

Exemplary results are shown in Section 4.

4. *Proof-of-concept* Experiments

Preliminary experiments have been conducted using approx. 30% of a popular UKB dataset (see Web-2). The set contains a large collection of similar images in groups of four (so that for any query from the dataset only three other images should be retrieved).

200 hundred randomly selected images have been used as queries. The BoW retrieval performances are summarized in Table 1 (the results are averaged for *haraff* and *hesaff* keypoints). The cutoff has been somehow arbitrarily set at 5 (i.e. slightly more than the *ground truth* number of returns, i.e. 3).

Table. 1. Mean Average Precision (mAP) for 5 top returns.

	SIFT – 64k words	SIFT – 4M words	CONSIFT – 1G words
mAP value	0.177	0.293	0.473

Fig. 2 shows an example of BoW retrieval for the variants listed in Table 1. Positive and negative examples of the subsequent verification by CONSIFT-based keypoint matching are given in Fig. 3.

Based in the conducted experiments (the reported results are just a part of them) we can conclude that CONSIFT descriptors, in spite of much larger memory requirements, can be considered an attractive tool for a large scale image retrieval. Its main advantage is the capability to incorporate a certain amount of contextual (geometric and photometric) data into individual keypoint descriptors so that BoW-based retrieval can provide more credible results and, secondly, the subsequent verification of preliminary selected images can be performed at the level of individual keypoint matching. Thus, the most time consuming element of retrieval in large-scale visual databases can be prospectively abandoned.

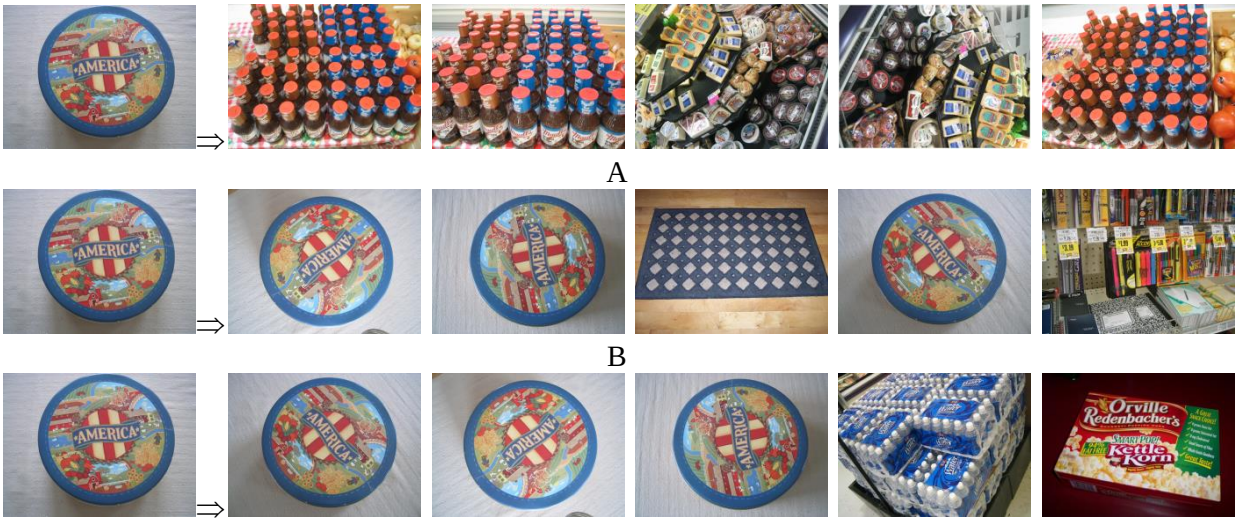

Fig. 2. Top retrievals by BoW using (A) SIFT (64K words), (B) SIFT (4M words) and (C) CONSIFT (1G words).

Fig. 3. Positive and negative verification of Fig. 2C retrievals by multi-CONSIFT keypoint matching (P = 3).

References

Chum O., Perdoch M., Matas J. (2009). Geometric min-hashing: Finding a (thick) needle in a haystack. Proc. IEEE Conf. CVPR'09, 17–24.

Chum O., Matas J. (2010). Large-scale discovery of spatially related images. IEEE Trans. PAMI 32(2), 371-377.

Csurka G., Dance C., Fan L.X., Willamowski J., Bray C. (2004). Visual categorization with bags of keypoints. Proc. ECCV'2004 Workshop on Statistical Learning in Computer Vision.

Jegou H., Douze M., Schmid C. (2010). Improving bag-of-features for large scale image search. Int. J. Comp. Vision 87(3), 316–336.

Lowe D. (2004). Distinctive image features from scale-invariant keypoints. Int. J. Com. Vision 60(1), 91–110.

Nister D., Stewenius H. (2006). Scalable recognition with a vocabulary tree. Proc. IEEE Conf. CVPR'06, 2161-2168.

Sluzek A. (2014). Contextual descriptors improving credibility of keypoint matching:Harris-Affine, Hessian-Affine and SIFT feasibility study. Submitted to ICARCV 2014.

Stewenius H., Gunderson S.H., Pilet J. (2012). Size matters: Exhaustive geometric verification for image retrieval. LNCS 7573 (Proc. ECCV 2012), 674-687.

Tolias M., Jegou H. (2014). Visual query expansion with or without geometry: refining local descriptors by feature aggregation. Pattern Recognition (accepted).

Web sites:

Web-1: http://www.robots.ox.ac.uk/~vgg/research/affine/, consulted 19 May, 2014.

Web-2: http://vis.uky.edu/~stewe/ukbench/, consulted 19 May, 2014.