

Partial Near-Duplicate Detection in Random Images by a Combination of Detectors

Andrzej Śluzek

Khalifa University, Abu Dhabi, UAE
andrzej.sluzek@kustar.ac.ae

Abstract. Detection of partial near-duplicates (e.g. similar objects) in random images continues to be a challenging problem. In particular, scalability of existing methods is limited because keypoint correspondences have to be confirmed by the configuration analysis for groups of matched keypoints. We propose a novel approach where pairs of images containing partial near-duplicates are retrieved if ANY number of keypoint matches is found between both images (keypoint descriptions are augmented by some geometric characteristics of keypoint neighborhoods). However, two keypoint detectors (Harris-Affine and Hessian-Affine) are independently applied, and only results confirmed by both detectors are eventually accepted. Additionally, relative locations of keypoint correspondences retrieved by both detectors are analyzed and (if needed) outlines of the partial near-duplicates can be extracted using a keypoint-based co-segmentation algorithm. Altogether, the approach has a very low complexity (i.e. it is scalable to large databases) and provides satisfactory performances. Most importantly, *precision* is very high, while *recall* (determined primarily by the selected keypoint description and matching approaches) remains at acceptable level.

Keywords: keypoint description, keypoint correspondences, partial near-duplicates, affine invariance, object detection, co-segmentation.

1 Introduction and Background Work

Detection of partial near-duplicates (e.g. retrieval of image pairs containing the same objects on diversified backgrounds) is a challenging problem for which a fully scalable solution has not been found yet. Because individual keypoint matches are usually incorrect in a (semi-)global context, post-processing operations have to be performed, where the spatial consistency over groups of preliminarily matched keypoints is verified. This is a computation-intensive task, no matter whether the Hough transform (e.g. [6], [9]), RANSAC-based methods (e.g. [1], [17]) or other less popular solutions (e.g. [19]) are used.

Currently, most of the *state-of-the-art* methods (e.g. [3], [5], [2]) seem to apply this approach, although they attempt to preliminarily reduce the numbers of analyzed image pairs using, for example, (*geometric*)*min-hashing* or *weak geometric*

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: [10.1007/978-3-319-02895-8_64](https://doi.org/10.1007/978-3-319-02895-8_64)

consistency. Nevertheless, with such approaches the size of visual databases cannot grow indiscriminately. In particular, there is always a need to process groups of matched keypoints in all pairs of preselected images (geometric verification).

In this paper, we attempt to solve the problem of partial near-duplicate detection using only individual keypoint matches. The basic idea is to incorporate into descriptions of individual keypoints selected visual and geometric characteristics of keypoint neighborhoods. Similar concepts of *keypoint bundling* have been discussed previously (e.g. [17], [10] and [12]). However, in most cases keypoint bundles are used as a pre-retrieval mechanism, i.e. matched bundles indicate for which image pairs (and at which locations within these images) geometric consistency of matched keypoints should be verified. Only in [12] keypoint bundles are represented by affine-invariant descriptions which are incorporated into descriptors of keypoints around which the bundles are built (such keypoints are referred to as bundle centroids). Then, a match between two bundles indicates that there is some photometric *and* geometric similarity between two groups of keypoints (incorporated into both bundles) so that the presence of partial near-duplicates can be assumed without any further geometric verification. This method, when using vocabularies of reasonable size to represent image contents and geometry, provides acceptable *precision* and *recall* (both reaching approx. 50% level, details in [12]).

We apply a very similar approach, i.e. keypoint description incorporating visual and geometric characteristics of keypoint neighborhoods. However, compared to [12], three significant changes have been introduced:

- (a) The geometric model of keypoint bundles is simplified (in order to accept stronger distortions). On one hand, it improves *recall* of partial near-duplicate retrieval, but on another hand *precision* deteriorates.
- (b) Two variants of the method using alternative types of keypoint detectors (Harris-Affine and Hessian-Affine, see [7]) are run simultaneously, and only pairs of images retrieved by both variants are preliminarily accepted. Thus, a high level of *recall* is maintained, while *precision* is much higher than achieved by individually applied variants.
- (c) Finally, keypoint correspondences are accepted if similarly located keypoint correspondences exist for the other detector. This step further improves *precision*, which reaches nearly 100%.

Principles of keypoint description (both the previous version and the proposed improvements) are highlighted in Section 2. In Section 3, we describe details of partial near-duplicate retrieval by using a combination of Harris-Affine and Hessian-Affine results (including the post-processing operations mentioned in the above Step 3).

Section 4 presents exemplary verification results for the selected datasets. Finally, Section 5 concludes the paper and highlights the directions for current and future researches.

2 Keypoint Bundles

Assuming that keypoint matching is considered the main operation in partial near-duplicate retrieval, and accepting that individual *standard* (e.g. based on SIFT descriptors) keypoint correspondences are virtually useless in this problem (most of them are incorrect in (semi-)global image context, e.g. [10]) we propose to incorporate characteristics of keypoint neighborhoods into keypoint description. Obviously, neighborhoods of limited size (either the radius or the number of neighboring keypoints) should be used. However, we exclude from the neighborhoods keypoints which are too close to the center or are significantly smaller/larger than the central keypoint. Altogether (as shown in Fig.1a) given a keypoint K represented by E ellipse, its neighborhood contains other keypoints K_i (with E_i ellipses) for which the following conditions are satisfied:

1. The Mahalanobis distance $D_M(K, K_i)$ is between $0.5\sqrt{2}$ and 2 (where the unit distance is defined by the shape of E ellipse).
2. The area of E_i ellipse is between 0.5 and 1.5 of the area of E ellipse..

Additionally, if more that 20 keypoints fulfill Conditions 1 and 2, only 20 of them (the closest to K) are retained so that the computational complexity of neighborhood processing is constrained.

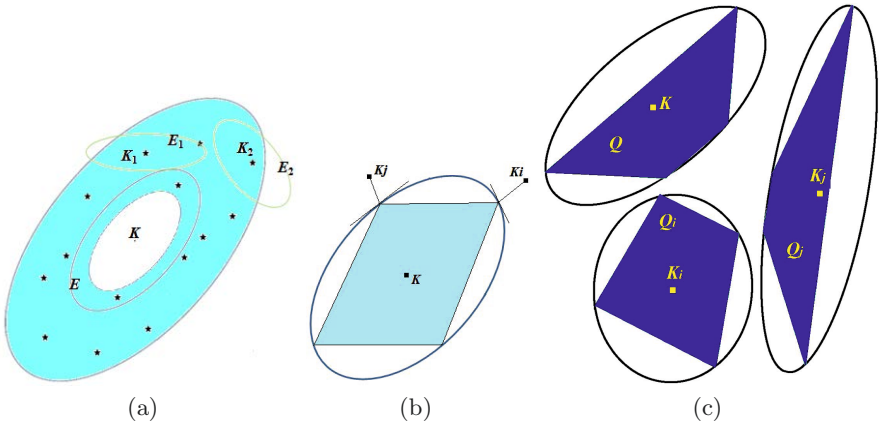


Fig. 1. Exemplary distribution of keypoints in the neighborhood of K keypoint (a). A trapezoid built in E ellipse in the context of K_i and K_j keypoints (b), and the trapezoids built in all three ellipses of a tuple (c).

Subsequently, all pairs of the neighborhood keypoints K_i and K_j (if they also are not too close to one another, see Condition 1 above) are used to form tuples $\{K, K_i, K_j\}$. The collection of such tuples is referred to as the *keypoint bundle* with K centroid. In practice, the average number of tuples in a bundle for typical images of fairly complex contents is approx. 60 – 70 for both Harris-Affine and Hessian-Affine keypoints.

2.1 Description of Keypoint Bundles

Photometric characteristics of an individual tuple of keypoints can be invariantly represented by the corresponding SIFT descriptors (or SIFT words) of K , K_i and K_j keypoints. The (affine-)invariant representation of the tuple's geometry is more complicated. It has been shown in [12,14] how several shapes can be unambiguously built within a tuple $\{K, K_i, K_j\}$. We use only some of these shapes, namely the trapezoids found in ellipses in the context of two other keypoints. Fig.1b illustrates (more details in [14]) how a trapezoid is built inside E ellipse (centered in K keypoint), while Fig.1c shows the trapezoids Q , Q_i and Q_j correspondingly built for all keypoints of the tuple.

Since the shapes of such trapezoids change co-variantly with any affine mapping of the tuple, we use (following [12,14]) the simplest affine-invariant moment-based shape descriptor Inv (Eq. 2, details in [15]) computed over the three trapezoids to affine-invariantly represent the configuration of the tuple. Therefore, geometric characteristics of each tuple are described by a 3D vector

$$[Inv(Q), Inv(Q_i), Inv(Q_j)], \quad (1)$$

where

$$Inv = \frac{M_{20}M_{02} - M_{11}^2}{M_{00}^4} \quad (2)$$

(note that M_{pq} is the central moment of order $p + q$).

Altogether, the tuple is described photometrically and geometrically by a 6D vector

$$[Sift(K), Sift(K_i), Sift(K_j), Inv(Q), Inv(Q_i), Inv(Q_j)]. \quad (3)$$

Then, the whole bundle centered at K keypoint is represented by a list of such vectors (one for each tuple in a bundle). In practice, 5D vectors

$$[Sift(K_i), Sift(K_j), Inv(Q), Inv(Q_i), Inv(Q_j)] \quad (4)$$

can be used because $Sift(K)$ is the same for all tuples and it can be memorized only once.

3 Detection of Partial Near-Duplicates

3.1 Matching Keypoint Bundles

The proposed descriptions of keypoint bundles actually represent semi-local structures of images (i.e. keypoints with their neighborhoods). Thus, a match between bundles around two keypoints indicates similarity between image fragments much larger than individual keypoint ellipses. In other words, this is an indicator of partial near-duplicates in both images.

For convenience, the matching operation for two bundles built around keypoints K and L , correspondingly, is divided into two phases. First, $Sift(K)$ and $Sift(L)$ are compared (i.e. we match the bundle centroids). Any typical approach can be used, e.g. *mutual-nearest-neighbor* or *the-same-visual-word*. In the conducted experiments, a SIFT vocabulary of 2^{16} words has been used.

If K and L match, their bundles are compared by matching tuples from both bundles. Finally, we assume that the bundles match, if at least A matching tuples are found for which

$$\begin{aligned} & [Sift(K_i), Sift(K_j), Inv(QK), Inv(QK_i), Inv(QK_j)] \equiv \\ & \equiv [Sift(L_m), Sift(L_n), Inv(QL), Inv(QL_m), Inv(QL_n)]. \end{aligned} \quad (5)$$

We match tuples by *the-same-word* approach, where SIFT descriptors are quantized into a relatively small vocabulary of 2000 words, while Inv invariants are quantized into 12 words only. Note that the tuple geometry is represented by three values only (compared to 16D vectors in [12]) which are quantized very coarsely so that a wide range of geometric image deformations can be tolerated.

The number of matching tuples (A threshold) needed for a match between two bundles has been established experimentally. It has been finally decided to use $A = 2$ for both Harris-Affine and Hessian-Affine bundles (even though the former ones usually contain slightly more tuples).

3.2 Preliminary and Final Image Matching

Detection, bundling and matching operations are performed independently using Harris-Affine and Hessian-Affine keypoints. When compared images contain clearly visible partial near-duplicates, usually both approaches retrieve a number of matching bundle pairs (i.e. correspondences between bundle centroids and unspecified numbers of similar tuples) as shown in a simple example in Figs 2a and 2b. However, the number of such matches is hard to predict, and even images of random contents may be occasionally matched as well (although a close visual inspection always reveals some level similarity between the corresponding areas). Nevertheless, such correspondences are rather infrequent in random images and (in general) Harris-Affine and Hessian-Affine matches are found at different locations, as illustrated in Figs 2c and 2d. In images sharing actual partial near-duplicates, however, Harris-Affine and Hessian-Affine matches are usually located in the same areas (as seen in Figs 2a and 2b).

Since the numbers of matching keypoints (bundles) are unpredictable, we preliminarily assume that pairs of images with *any* number of matches may contain partial near-duplicates. However, if it is additionally requested that *both* Harris-Affine and Hessian-Affine matches (see Point (b) in Section 1) must exist in the image pair, the number of false correspondences is dramatically reduced.

A further reduction (see Point (c) in Section 1) is obtained by checking the locations of matched keypoints (bundle centroids). A pair of images is retained

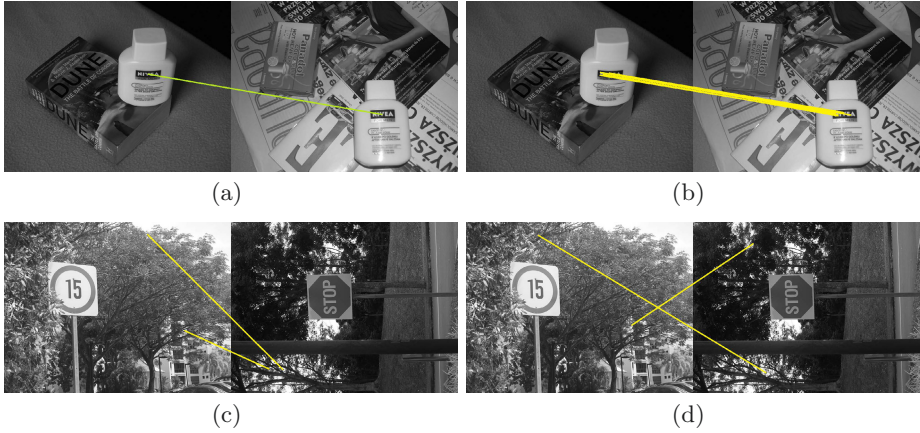


Fig. 2. Matched keypoints (centroids of matched bundles) for images sharing the same object (using Hessian-Affine (a) and Harris-Affine (b)). The results for images of random contents are shown in (c) and (d).

if each pair of matched Harris-Affine keypoints (K_{har}, L_{har}) has at least one counterpart pair of matched Hessian-Affine keypoints (K_{hes}, L_{hes}) (and another way around) with similar coordinates, i.e.

$$K_{hes} \in E(K_{har}); L_{hes} \in E(L_{har}); K_{har} \in E(K_{hes}); L_{har} \in E(L_{hes}) \quad (6)$$

where $E(K)$ and $E(L)$ are ellipses of the corresponding keypoints .

It can be seen that matched keypoints in Figs 2a and 2b clearly satisfy Eq.6, while the pairs of images from Figs 2c and 2d would be rejected.

The proposed method of detecting images with partial near-duplicates is very fast and efficient. Although matching using two types of keypoints is needed (i.e. the database memory for image representation is doubled), no geometric verification of keypoint matches (which is the bottleneck of existing solutions) is needed. Although some geometry-based calculations are performed in Eq.6, their complexity is negligible.

Experimental verification of this proposed method is presented in the following section.

4 Experimental Verification

4.1 Methodology

The experiment has been conducted using two publicly available datasets, i.e. VISIBLE and PASCAL 2007. VISIBLE¹ contains diversified views of 1, 2 or 3

¹ <http://156.17.10.3/~visible/data/upload/FragmentMatchingDB.zip>

locally planar objects on diversified backgrounds. The objects are manually outlined so that *ground-truth* (the presence of partial near-duplicates) is estimated. Actually, other partial near-duplicates (outside the object outlines) also exist (see examples below) so that this is a very conservative *ground-truth*. PASCAL 2007² also provides *ground-truth* data but they are not partial near-duplicates (instead, they are outlines of the same category objects, which may look very differently) so that we consider this dataset a collection of confusing images. Therefore, we assume only 511 *ground-truth* image pairs with partial near-duplicates (the number of VISIBLE image pairs sharing the same object(s)). The total number of image pairs is 4,950 in VISIBLE only (these are used in the first part of the experiment) and 135,460 in the union of VISIBLE and PASCAL 2007 (the second part of the experiment). Images in Fig.2 are actually from VISIBLE dataset.

Bundle centroids are matched by using thresholded difference between SIFT descriptors (the threshold obtained from over 50,000,000 *mutual-nearest-neighbor* matches). Neighborhood keypoints are matched using a 2000 word SIFT vocabulary, while the tuple geometries are compared using the vocabulary of $12^3 = 1728$ words (*Inv* invariant quantization in Section 3.1).

4.2 Results

Full results obtained for VISIBLE dataset (i.e. matches between 4,950 image pairs attempted) are summarized in Table 1, and exemplary correct matches (both Harris-Affine and Hessian-Affine) are shown in Fig. 3. Note that matches in Fig.2 are also from this experiment.

Table 1. Retrieved image pairs (total and correct, compared to *ground truth*) in VISIBLE dataset

Method	Total	Correct	Precision	Recall
HarAff	536	306	57.09%	59.88%
HesAff	488	284	58.20%	55.58%
HarAff +HesAff	375	283	75.47%	55.38%
HarAff+ HesAff+Eq.6	304	283	93.09%	55.38%

Recall of the ultimate results is not perfect, but still better (55.38% versus 51.40%) than reported in [12] for the same dataset. *Precision*, however, is very high and it effectively reaches almost 100%. This can be claimed because most of false positives are actually correct (indicating near-duplicate fragments outside the *ground truth* outlines of objects). Examples of such *correct false positives* are provided in Fig. 4.

² <http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/>

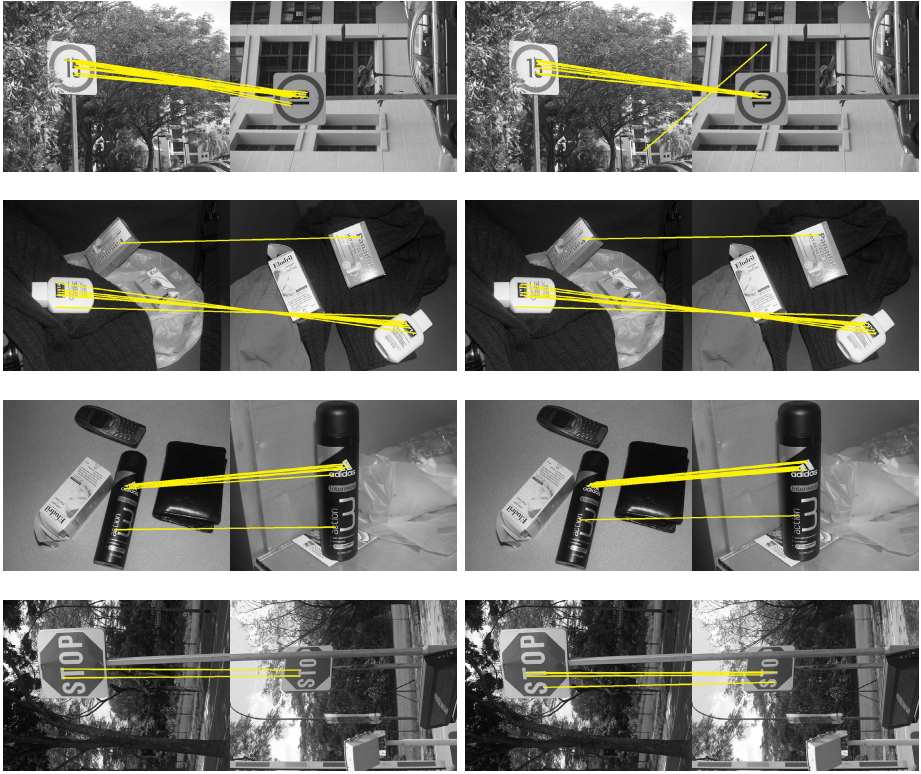


Fig. 3. Matched centroids of bundles for VISIBLE images sharing the same object. Hessian-Affine (left column) and Harris-Affine (right column) detectors are used.

A limited *recall* value can be attributed to certain effects in the keypoint bundling process. Our experiments show that 35-40% of detected keypoints (both Harris-Affine and Hessian-Affine) have too few neighbors (as defined in Section 2, see Fig.1a) to form bundles with a sufficient (for prospective bundle matching) number of tuples. If matches between such keypoints are the only evidences of partial near-duplicity between two image fragments, those partial near-duplicates would be missed.



Fig. 4. Examples of *correct false positives*, i.e. near-duplicate fragments identified outside the *ground truth* objects

Results for the union of VISIBLE and PASCAL 2007 datasets (with 135,460 image pairs to be matched) are presented in Table 2. It can be noticed that when only Harris-Affine keypoint bundles are matched *precision* is much lower than in the first experiment. This is understandable because stray partial near-duplicates (i.e. fragments with weakly seen visual similarity usually represented by only one match in the whole image) appear quite often, i.e. in approx. 2.5% of image pairs. However, such random matches usually happen for only Harris-Affine or Hessian-Affine keypoint. Thus, as shown in the table, the intersection of Harris-Affine and Hessian-Affine retrievals provides much higher *precision*. Eventually, after the verification by Eq. 6, *precision* is almost the same as in the first experiment (where the number of image pairs is $27\times$ smaller).

Table 2. Retrieved image pairs (total and correct, compared to *ground truth*) in VISIBLE and PASCAL 2007 datasets

Method	Total	Correct	Precision	Recall
HarAff	3,453	306	8.86%	59.88%
HarAff +HesAff	660	283	42.88%	55.38%
HarAff+ HesAff+Eq.6	318	283	88.99%	55.38%

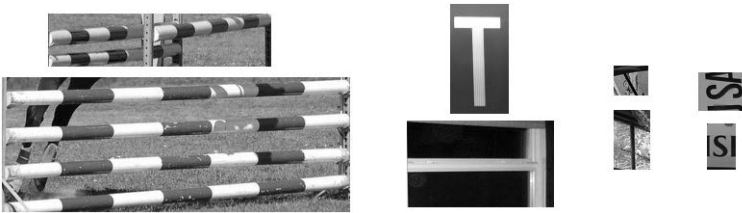


Fig. 5. Examples of near-duplicate fragments identified outside the *ground truth* objects (VISIBLE + PASCAL 2007 datasets)

Similarly to Table 1, the actual *precision* in the last row of Table 2 is also almost 100%. Fig. 5 shows examples of fragments which are clearly partial near-duplicates, but which are not included into the *ground truth* (thus, considered *false positives*).

4.3 Additional Operations

The present method of partial near-duplicate retrieval returns only image pairs containing near-duplicate fragments and provides approximate locations of these

fragments using coordinates of matched keypoints (i.e. centroids of matched bundles) in both images. If the outlines of near-duplicates are required, additional operations should be performed. Details of such operations are not discussed in this paper, but their exemplary outcomes are presented for completeness.

Although outlines of partial near-duplicates can be approximated by convex polygons using the methods proposed in [9], we prefer another technique based on the concept of *co-segmentation*.

Popular co-segmentation methods (e.g. [4,8]) use the graph-cut framework solved by minimizing a Markov Random Field energy function through a min-cut/max-flow algorithm. The method we apply has been adopted from an unpublished report [18]. This algorithm follows the standard approaches regarding the image energy (which consists of the *deviation penalty* and *separation penalty* functions). However, a novel foreground energy is proposed based in non-linear mappings between co-segmented images. The mappings (based on TPS, i.e. *thin plate splines* warping) use the keypoint correspondences established in partial near-duplicate detection as the control points. In the report, the algorithm is benchmarked against alternative solutions, and its performances in co-segmentations of partial near-duplicates have been found superior to other methods. Fig.6 shows exemplary image fragments around matched keypoints, and the results of co-segmentation.

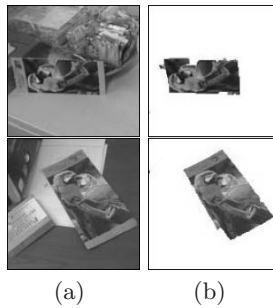


Fig. 6. Exemplary pair of approximately matched image fragments (a) and the co-segmentation results using the [18] algorithm

5 Summary

5.1 Discussion

The paper presents a novel method for detection of partial near-duplicates in large databases of unknown and unpredictable images. Using keypoint detection as a starting point, we build around detected keypoints *bundles of neighboring keypoints*. Bundles are described by 5D vectors invariantly representing photometric and geometric properties of the bundles.

The bundle descriptions are incorporated into descriptors of keypoints (bundle centroids) so that individual keypoint correspondences (found using such augmented descriptors) indicate without any geometric verification that images may contain partial near-duplicates around the locations of matched keypoints.

Performances of the method (*precision* in particular) are improved by intersecting results obtained by two independently applied affine-invariant keypoint detectors, i.e. Harris-Affine and Hessian-Affine.

The proposed description of keypoint bundles is effectively a *set-of-words* from a large vocabulary. If bundle centroids are matched using a SIFT vocabulary of 2^{16} words, while tuples are matched using two words from a 2000-word SIFT vocabulary and a vocabulary of 1728 words to represent geometry (see Section 4.1), the overall size of the vocabulary is more than 4.5×10^{14} . With such a huge vocabulary, sophisticated image indexing strategies and/or efficiently organized databases (e.g. [11,16]) can be implemented for prospective web-scale applications of the method.

A particularly attractive properties of the proposed approach is that, in spite of a huge vocabulary, a very good balance is maintained between *precision* and *recall*. Usually (see a discussion in [16]) too large vocabularies are unable to produce satisfactory *recalls*. In our method, the value of *recall* is acceptable and, actually, it can be further improved by modifying parameters for neighborhood building and bundling keypoints (see Section 4.2).

5.2 Future Works

The presented method can be considered fully developed in terms of its methodological principles. However, numerous technical improvements are possible. In particular, the method is currently implemented in Matlab so that we do not discuss its timing performances. They will be experimentally verified after an efficient C++ implementation (incorporating additional mechanisms, e.g. inverted indexing and distributed memory for inverted files, [13]) will have been developed.

Moreover, extensive experiments on much larger dataset sets are needed for fine-tuning parameters (e.g. the size and shape of keypoint neighborhoods, threshold values, etc.) and general evaluation. Many ideas will be borrowed from a recent Google project preliminarily presented in [16].

References

1. Chum, O., Matas, J.: Matching with prosac - progressive sample consensus. In: Proc. IEEE Conf. CVPR 2005, San Diego, CA, pp. 220–226 (2005)
2. Chum, O., Matas, J.: Large-scale discovery of spatially related images. IEEE PAMI 32(2), 371–377 (2010)
3. Chum, O., Perdoch, M., Matas, J.: Geometric min-hashing: Finding a (thick) needle in a haystack. In: Proc. IEEE Conf. CVPR 2009, pp. 17–24 (2009)
4. Hochbaum, D., Singh, V.: An efficient algorithm for co-segmentation. In: Proc. ICCV 2009, Kyoto, pp. 269–276 (2009)

5. Jegou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. *International Journal of Computer Vision* 87(3), 316–336 (2010)
6. Lowe, D.G.: Object recognition from local scale-invariant features. In: *Proc. 7th IEEE Int. Conf. Computer Vision*, vol. 2, pp. 1150–1157 (1999)
7. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. *International Journal of Computer Vision* 60, 63–86 (2004)
8. Mukherjee, L., Singh, V., Dyer, C.R.: Half-integrality based algorithms for cosegmentation of images. In: *Proc. IEEE Conf. CVPR 2009*, Miami Beach, pp. 2028–2035 (2009)
9. Paradowski, M., Śluzek, A.: Local keypoints and global affine geometry: Triangles and ellipses for image fragment matching. In: Kwaśnicka, H., Jain, L.C. (eds.) *Innovations in Intelligent Image Analysis*. SCI, vol. 339, pp. 195–224. Springer, Heidelberg (2011)
10. Romberg, S., August, M., Ries, C.X., Lienhart, R.: Robust feature bundling. In: Lin, W., Xu, D., Ho, A., Wu, J., He, Y., Cai, J., Kankanhalli, M., Sun, M.-T. (eds.) *PCM 2012*. LNCS, vol. 7674, pp. 45–56. Springer, Heidelberg (2012)
11. Sivic, J., Zisserman, A.: Efficient visual search of videos cast as text retrieval. *IEEE PAMI* 31(4), 591–606 (2009)
12. Śluzek, A.: Large vocabularies for keypoint-based representation and matching of image patches. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) *ECCV 2012 Ws/Demos*, Part I. LNCS, vol. 7583, pp. 229–238. Springer, Heidelberg (2012)
13. Śluzek, A.: Inverted indexing in image fragment retrieval using huge keypoint-based vocabularies. In: *Proc. CBMI 2013*, Veszprem, pp. 167–172 (2013)
14. Śluzek, A., Paradowski, M.: Detection of near-duplicate patches in random images using keypoint-based features. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P., Zemčik, P. (eds.) *ACIVS 2012*. LNCS, vol. 7517, pp. 301–312. Springer, Heidelberg (2012)
15. Śluzek, A.: *Zastosowanie metod momentowych do identyfikacji obiektów w cyfrowych systemach wizyjnych*. WPW, Warszawa (1990)
16. Stewénius, H., Gunderson, S.H., Pilet, J.: Size matters: Exhaustive geometric verification for image retrieval. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012*, Part II. LNCS, vol. 7573, pp. 674–687. Springer, Heidelberg (2012)
17. Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling features for large scale partial-duplicate web image search. In: *Proc. IEEE Conf. CVPR 2009*, Miami Beach, pp. 25–32 (2009)
18. Yang, D., Śluzek, A.: Co-segmentation by keypoint matching: Incorporating pixel-to-pixel mapping into mrf. *Tech. rep.*, Nanyang Technological University, SCE, Singapore (2010)
19. Zhao, W.-L., Ngo, C.-W.: Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection. *IEEE Trans. on Image Processing* 2, 412–423 (2009)