Extended Keypoint Description and the Corresponding Improvements in Image Retrieval

Andrzej Śluzek

Khalifa University, Abu Dhabi, United Arab Emirates

Abstract. The paper evaluates an alternative approach to BoW-based image retrieval in large databases. The major improvements are in the re-ranking step (verification of candidates returned by BoW). We propose a novel keypoint description which allows the verification based only on individual keypoint matching (no spatial consistency over groups of matched keypoints is tested). Standard Harris-Affine and Hessian-Affine keypoint detectors with SIFT descriptor are used. The proposed description assigns to each keypoint several words representing photometry and geometry of the keypoint in the context of neighbouring image fragments. The words are Cartesian products of typical SIFT-based words so that huge vocabularies can be built. The preliminary experiments on several popular datasets show significant improvements in the pre-retrieval phase combined with a dramatically lower complexity of the re-ranking process. Because of that, the proposed methodology is particularly recommended for the retrieval in very large datasets.

1 Introduction

Keypoint-based image matching is one of the fundamental tools in CBVIR. Even though the reported solutions differ in keypoint detectors, keypoint descriptors and the vocabulary sizes (matching using the original descriptor vectors is computationally inefficient) typical approaches to the retrieval of similar images or sub-images generally follow the same two-step scheme. First, the candidate images are pre-retrieved. One of the standard techniques is BoW (e.g. [1]) where the sparse histograms of visual words are matched to find similar images. This model ignores the spatial distributions of keypoints so that the second step of geometric/configurational verification is needed to re-rank the pre-retrieved candidates, i.e. to identify the most similar images (or similar fragments within them) from the pool of candidates. Computational complexity of the second step is high, and many attempts have been reported (e.g. [2–4], etc.) to simplify it. Nevertheless, the retrieval algorithms cannot be considered sufficiently scalable as long as spatial distributions of matching keypoints have to be analyzed.

In this paper, we evaluate an alternative approach where the first level of BoW-based pre-retrieval is retained, but the complexity of the second level is dramatically reduced. This approach is based on the concept of *contextual keypoint descriptors* (preliminarily introduced in [5]) which are built using dependencies between keypoints extracted by two complementary detectors, e.g. Harris-Affine and Hessian-Affine. Each keypoint is represented by several descriptors, i.e. by several words, so that keypoint matching is more flexible (the level of similarity is defined by the number of words shared by two descriptions). Another advantage of such extended descriptions is that similarities between larger image fragments can be established using only matches between individual keypoints (no spatial analysis needed!). It is believed the presented approach may contribute to development of fully scalable CBVIR algorithms.

In Section 2, the background works are briefly overviewed, in particular the works related to the proposed description. Section 3 explains (and illustrates on selected examples) advantages of the method, which is experimentally verified using several popular datasets. Concluding remarks and observations are included in Section 4.

2 Background works

2.1 Keypoint matching

Performances of keypoint matching depend on the quality of keypoint detectors (this aspect is not discussed in the paper) and on credibility of the matching scheme. The *mutual nearest neighbour* O2O scheme is generally considered (see [6]) the most credible one so that we use its results as the benchmark to evaluate matching based on visual words. Two standard detectors, i.e. Harris-Affine (*haraff*) and Hessian-Affine (*hesaff*) [7] are selected (the reasons for this choice are later explained in detail) and SIFT [8] is the selected keypoint descriptor because of its popularity and high repeatability. Actually, we use its RootSIFT variant whis was reported superior in [9]. Performances of keypoint matching are evaluated on a popular benchmark dataset of diversified images[1]. The dataset provides homographies between *the-same-category* images, so that the ground truth keypoint correspondences can be identified similarly to [10].

Table 1 summarizes the results of keypoint matching using both O2O and visual vocabularies of diversified sizes. The ranges of values have been obtained by using several alternative i.e. generated from different populations of images) vocabularies of each size and/or using sets of keypoints extracted by two detectors. It should be also highlighted that (unlike in most work using the same benchmark dataset) *all* pairs of images are compared to better reflect scenarios of larger-scale image retrieval.

Although satisfactory *recall* can be achieved by using small vocabularies, low *precision* values (which further deteriorate if more images are added to the dataset) confirm a well-known fact that credible image retrieval based only on individual keypoint matches is not reliable. It can be also noticed that the overall performance of keypoint matching (represented by *F-measure*) improves with the size of vocabulary, but this size cannot be indiscriminately increased. If a vocabulary grows too large, the quantization intervals become smaller that the

[1] http://www.robots.ox.ac.uk/ vgg/research/affine/

Table 1. Performances of keypoint matching using *haraff* and *hesaff* keypoints (see also [5]).

Measure	O2O	2^{10} words	2^{16} words	2^{20} words	2^{25} words	2^{30} words
Recall	0.571-0.587	0.584-0.608	0.266-0.281	0.209-0.219	0.173-0.176	0.141-0.142
Precision	0.104-0.114	0.002-0.003	0.003-0.004	0.035-0.076	0.071-0.124	0.132-0.178
F-measure	0.177-0.190	0.003-0.005	0.005-0.008	0.060-0.112	0.080-0.144	0.137-0.158

natural fluctuations of descriptor values, and very few (if any) matches would be found even in pairs of highly similar images. Some sources (e.g. [11, 12]) indirectly indicate that several millions is the maximum practical size of visual vocabularies.

2.2 Extended descriptors

Descriptions of keypoints would be obviously enriched, if some data about the keypoint context can be incorporated. Intuitively, the context of a region-based feature can be defined as a collection of neighbouring contour-based features (and another way around). We propose, therefore, to use two complementary keypoint detectors (e.g. *hesaff* detecting blob-like features and *haraff* detecting corner-like features) and to combine their SIFT descriptors in a way explained in the definitions below and illustrated in Figs 1 and 2.

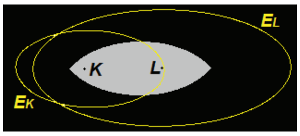


Fig. 1. Examples of a *haraff* keypoint (K) and a *hesaff* keypoint (L) extracted from a simple image.

Def.1 Given a *hesaff(haraff)* keypoint K and its neighbouring *haraff(hesaff)* keypoint L (with the corresponding ellipses E_K and E_L, see Fig. 1), the **CONSIFT** descriptor of K in the context of L is defined by a $384D$ vector which is a concatenation of three $128D$ SIFT descriptors: (a) the original SIFT computed over E_K ellipse, (b) SIFT computed over E_K ellipse with $\overrightarrow{K, L}$ vector as the reference orientation, and (c) SIFT computed over E_L ellipse with $\overrightarrow{L, K}$ vector as the reference orientation (see Fig. 2).

Thus, the first part of CONSIFT descriptors characterizes local properties of keypoints, while the remaining parts provide some data about photometric and geometric properties of keypoint neighbourhood.

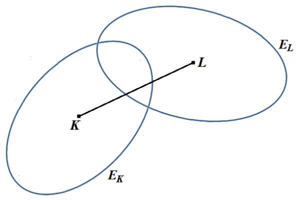


Fig. 2. A configuration of two keypoint ellipses for computing CONSIFT descriptor.

Def.2 Given a *hesaff(haraff)* keypoint K, its *extended description* consists of CONSIFT descriptors computed in the context of *haraff(hesaff)* L_i keypoints belonging to the neighbourhood of K. The neighbourhood is defined to follow a common-sense idea that a blob feature should be surrounded by a number of similar-scale corner features distributed approximately around the perimeter of the blob feature (or another way around).
Thus, L_i keypoints are considered neighbourhood keypoints if:

1. The Mahalanobis distances D_M between K and L_i satisfy:

$$1/\sqrt{2} \leq D_M \leq 2, \tag{1}$$

   where the unit distance is defined by the shape of E_K ellipse.
2. The areas of E_K and E_i ellipses are similar (i.e. the ratio is between 0.5 and 2).

Using a large set of test images, we have verified that the average size of such neighbourhoods is 8-10 (both for *haraff* and *hesaff* keypoints). If necessary, the maximum size can be constrained (e.g. not more than 20).

The extended descriptions are not particularly suitable for a direct keypoint matching (because of a high dimensionality of CONSIFT vectors). However, they can be conveniently used for matching by visual words. Each CONSIFT descriptor is actually a union of three SIFTs. Therefore, CONSIFT vocabularies can be built as Cartesian products of the original SIFT vocabularies. Even if those SIFT vocabularies are relatively small (i.e. the quantization is coarse) the resulting CONSIFT vocabularies are huge. For example, 1000-word SIFT vocabularies generate a billion-word CONSIFT counterpart ($10^3 \times 10^3 \times 10^3 = 10^9$). Such a vocabulary is expected to combine high *precision* (a large number of words) with high *recall* (coarse quantization of the contributing words).

In the scheme based on extended descriptions two keypoints match, if their descriptions share at least one CONSIFT word. However, more flexible conditions can be easily defined, e.g. only keypoints sharing at least N (where $N > 1$) CONSIFT words in their extended descriptions are considered a match. This is a significant advantage over traditional word-based matching, where keypoints can share at most a single word.

Superficially, the proposed method may look similar to image matching by *visual phrases* (e.g. [13]) since both approaches consider keypoints in a wider context. However, visual phrases (as defined in [13]) need spatial analysis, first in the histogram building and later in the consistency verification phase. Thus, for the practicality over large databases the class of transformations consider in visual phrases (similarly to other methods incorporating geometric verification) is constrained, e.g. allowing only shifts and/or scale changes between matching images.

Higher performances of CONSIFT-based matching (using ether 1 or 2 two shared CONSIFT words) are shown in Table 2 on the same dataset. The performances are better even than the O2O scheme based on full SIFT vectors (compare to Table 1). Additionally, the table indicates that vocabularies with approx. 1 billion CONSIFT words (i.e. the underlying SIFT vocabularies have only 1000 words) have superior *F-measures* than 64-billion word vocabularies. It suggests that also the size of CONSIFT vocabularies cannot grow indiscriminately. However, for huge databases of images with significant numbers of keypoints, larger CONSIFT vocabularies are still recommended because they provide very high *precision* while the level of *recall* is less critical in such cases.

Table 2. Keypoint matching using extended descriptions of *haraff* and *hesaff* keypoints (N indicates the minimum number of shared CONSIFTs).

Measure	$2^{10+10+10}$ words ($N=1$)	$2^{10+10+10}$ words ($N=2$)	$2^{16+10+10}$ words ($N=1$)	$2^{16+10+10}$ words ($N=2$)
Recall	0.355-0.371	0.285-0.314	0.166-0.175	0.138-0.158
Precision	0.201-0.245	0.332-0.402	0.398-0.440	0.477-0.612
F-measure	0.265-0.290	0.322-0.334	0.241-0.243	0.225-0.237

It can be, therefore, claimed that the use of extended keypoint descriptions is justified even though the memory resources are significantly increased; a single keypoint has, in average, 8-10 CONSIFT words instead of a single SIFT word (some keypoints have multiple SIFT descriptors - see [8]).

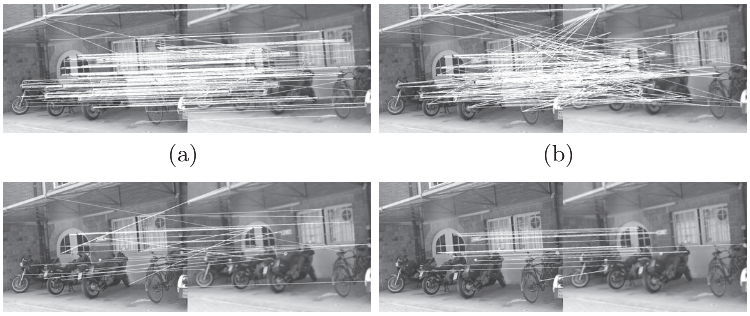
An illustrative example comparing keypoint matching by O2O scheme, SIFT words and CONSIFT words is given in Fig. 3.

3 Image retrieval

3.1 Bag-of-Words pre-retrieval

The ultimate objective of extended keypoint descriptions is to simplify image retrieval in very large databases. However, we do not intend to change the principles of BoW-based pre-retrieval returning images ranked by the BoW similarity.

In BoW representation (i.e. sparse histograms of word distributions in images) image similarities are approximated by the similarities between those histograms. Because our approach is proposed for databases of unknown and unpredictable sizes, the popular techniques of BoW normalization which require



(a) (b)

(c) (d)

Fig. 3. Keypoint matching results using: (a) O2O, (b) 2^{16} SIFT words, (c) 2^{25} SIFT words and (d) 2^{30} CONSIFT words.

database statistics (e.g. *td-idf*, [14]) cannot be applied, and we use histograms of *absolute* word frequencies in images.

Numerous measures of histogram similarities exist (e.g. [15]) but not all of them are applicable to BoW matching. Under the assumptions regarding BoW building in this work, we eventually selected a simple *histogram intersection* measure (proposed in [16]), where the distance between two histograms H_A and H_B over *Voc* vocabulary is defined by

$$d(H_A, H_B) = \sum_{w \in Voc} min(H_A(w), H_B(w)). \tag{2}$$

Such a measure nicely corresponds to the intuitive notion of similarity between both full images and sub-images (including textured images).

Additionally, to normalize the results over images with diversified numbers of keypoints, the Eq. 2 similarity between a query image A and a database image B is weighted by the factor S_F

$$S_F = 2^{(1 - n_B/n_A)}, \tag{3}$$

where n_A and n_B are the numbers of keypoints in the corresponding images. Such a normalization allows for more realistic pre-retrieval results in case of databases containing images with dramatically diversified numbers of keypoints (note that $S_F = 1$ for a pair of images with the same numbers of keypoints, $S_F < 1$ when the query has fewer keypoints, and $2 > S_F > 1$ for database images with fewer keypoints).

Performances of BoW pre-retrieval using SIFT and CONSIFT words have been tested on four popular datasets of very diversified characteristics, i.e. Oxford5k[2], UKB[3], Visible[4] and Caltech-Faces1999[5]. They represent diversified aspects of image retrieval, i.e. full images retrieval (UKB) or sub-image retrieval in easier (Oxford5k) and more complicated (Visible) scenarios, etc.

The results are shown in Table 3, which contains *mean average precision* (mAP) values obtained by using SIFT and CONSIFT vocabularies of several sizes. Because the objective is to evaluate performance variations (rather than the absolute values of mAP), the mAP values for the $64k$-word SIFT vocabulary are used as the reference (unit score). The presented scores are the average results for *haraff* and *hesaff* keypoints, and for several alternative vocabularies of each size.

Table 3. Relative *mean average precisions* (mAP) of image retrieval using SIFT and CONSIFT vocabularies of various sizes.

Dataset	$64k$ words SIFT	$1M$ words SIFT	$32M$ words SIFT	$1G$ words CONSIFT
Oxford5k	1.0	1.03	1.19	1.53
UKB	1.0	1.22	1.32	1.59
Visible	1.0	1.06	1.17	1.32
Faces1999	1.0	1.54	2.54	2.88

The content of Table 3 indicates that performances of BoW-based image pre-retrieval can be significantly improved if SIFT vocabularies are replaced by their CONSIFT counterparts. As an illustration, Figs 4-7 show the top rank returns by SIFT and CONSIFT for an exemplary queries from each tested dataset. We deliberately select not too successful examples to discuss improvements that can be subsequently introduced to in the second step (re-ranking pre-retrieved candidates).

3.2 Verification of pre-retrieved results

The content of Table 3 indicates that BoW pre-retrieval with CONSIFT words provides better performances (in terms of mAP values) than with SIFT words, but such results are not considered final. In other words, the verification step is still applied.

For the selected datasets, many works on consistency verification exist (e.g. [17] for Oxford5k, [4] for UKB and Oxford5k, [18] for Visible or [19] for Faces1999). They generally apply solutions which are computationally intensive (in spite of

[2] http://www.robots.ox.ac.uk/ vgg/data/oxbuildings/
[3] http://www.vis.uky.edu/ stewe/ukbench/
[4] http://156.17.10.3/ visible/data/upload/FragmentMatchingDB.zip
[5] http://www.vision.caltech.edu/html-files/archive.html

Fig. 4. Top rank BoW-returned images for an exemplary query in Visible dataset, using a CONSIFT vocabulary of $1G$ words (top row) and a SIFT vocabulary of $64k$ words (bottom row).

Fig. 5. Top rank BoW-returned images for an exemplary query in Oxford5k dataset, using a CONSIFT vocabulary of $1G$ words (top row) and a SIFT vocabulary of $64k$ words words (bottom row).

the reported simplification efforts). The only known example of a solution working without geometric verification, in [4], requires query expansion which should be considered computationally intensive as well.

We propose to reduce the verification step to a straightforward matching of extended keypoint descriptors, where a match between two keypoints is accepted if their descriptions share at least N (the recommended values of N are 3 or more) CONSIFT words (see Subsection 2.2). Such a mechanism effectively identifies pairs of keypoints which have sufficiently similar neighbourhoods. In other words, pre-retrieved images are accepted only if they contain larger fragments similar to some fragments of the query. In this mechanism, there is no difference between full and partial similarity of images (image retrieval *versus* sub-image retrieval) which in our opinion is another advantage of the method.

Fig. 6. Top rank BoW-returned images for an exemplary query in UKB dataset, using a CONSIFT vocabulary of $1G$ words (top row) and a SIFT vocabulary of $64k$ words (bottom row). Note that in UKB dataset each query has only three relevant returns, so that at least one return must be always incorrect.

Fig. 7. Top rank BoW-returned images for an exemplary query in Faces1999 dataset, using a CONSIFT vocabulary of $1G$ words (top row) and a SIFT vocabulary of $64k$ words (bottom row). Note that for the CONSIFT retrieval the incorrect face is actually placed on the same background as the query face.

Examples are provided in Fig. 8 which illustrates how selected images from Figs 4-7 are matched to the queries. Oviously, only images pre-retrieved by CONSIFT vocabularies are taken into account because for SIFT-based ranking it is generally impossible to distinguish between correct and incorrect pre-retrievals based on the spatial distributions of matching keypoints (an illustrative example is given in Fig. 9).

As shown in Fig. 8, the proposed method returns rather small numbers of keypoint correspondences pointing to the most similar fragments in both images. Images which are incorrectly pre-retrieved usually have no keypoint matches. If, however, some matches are found between the query and an (allegedly) incorrect image, those keypoint correspondences identify fragments which are, nevertheless, visually similar (although sometimes a careful inspection is needed to notice the actual existence of such a similarity).

The pre-retrieved images are subsequently re-ranked (similarly to the most popular solutions, e.g. [3, 17, 12, 4], etc.) based on the number of keypoint correspondences found. The experiments on the performances of re-ranked retrieval are still under way. Nevertheless, the preliminary results indicate that the per-

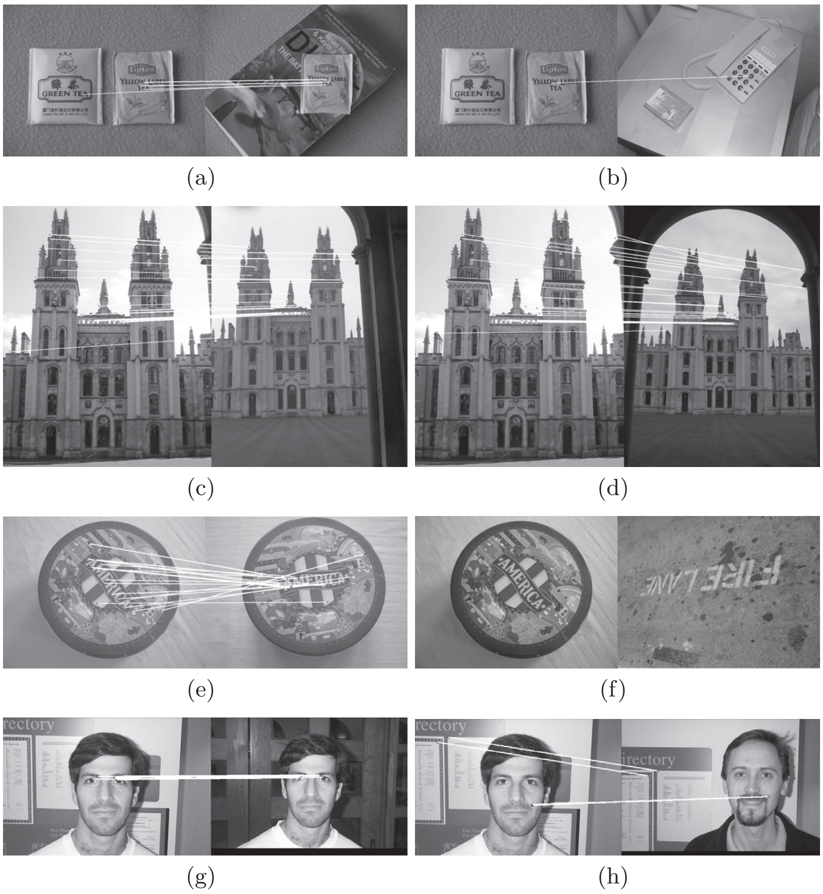
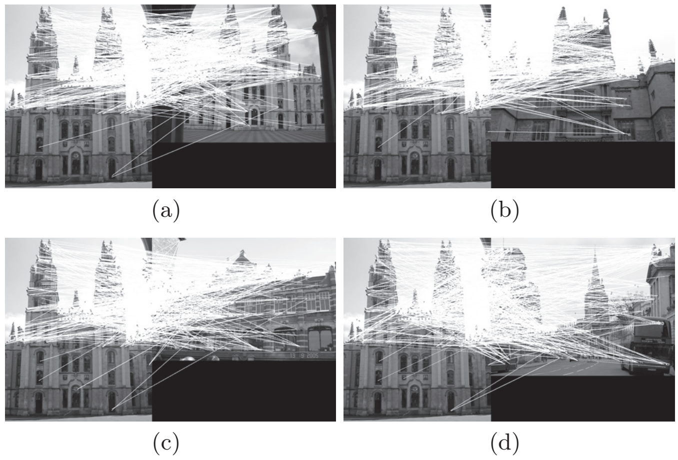


Fig. 8. Verification of BoW-pre-retrieved images (CONSIFT vocabulary) by matching extended descriptions of individual keypoints. Examples are from: (a,b) Visible dataset (Fig. 4), (c,d) Oxford5k dataset (Fig. 5), (e,f) UKB dataset (Fig. 6) and (g,h) Faces1999 dataset (Fig. 7).

formances are comparable to those reported in works using geometric verification. For example, the mAP improvements in the re-ranked lists over Oxford5k dataset are very similar to the improvements presented in [17].

Similarly to most works, we re-rank only a fixed number of top pre-retrievals (e.g. 300 threshold for Oxford5k dataset) even though more candidates are usually returned. However, with a huge size of CONSIFT vocabularies there are often cases when BoW pre-retrieval returns fewer images than the threshold number. This can be considered another advantage of the proposed approach (especially for large-scale applications).



(a) (b) (c) (d)

Fig. 9. Distributions of keypoint correspondences (SIFT words) in correctly (a) and incorrectly (b,c,d) pre-retrieved images (see Fig. 5).

4 Concluding remarks

The paper presents the fundamentals and preliminary experimental results of a novel BoW-based approach to image retrieval. Instead of typical vocabularies (e.g. derived from SIFT descriptors on *haraff*/*hesaff* keypoints) we use vocabularies built from CONSIFT descriptors to represent each keypoint by *extended descriptions* consisting of several CONSIFT-based words.

CONSIFT descriptors are concatenations of three SIFT vectors computed over a pair of *hesaff-haraff* keypoints within a predefined neighbourhood of a keypoint. Eventually, each keypoint is represented by several words (8-10 in average) so that keypoint matching is more flexible (the level of similarity can be estimated by the number of CONSIFT words shared by two descriptions).

A standard two-level model of retrieval is assumed, i.e. the BoW-based pre-retrieval of candidate images is followed by the verification of configuration constraints in the candidate images. No changes are introduced to the first level. Nevertheless, it has been preliminarily verified on popular datasets that performances of BoW pre-retrieval are improved if CONSIFT words are used. The major improvement, however, is at the second level. By using CONSIFT words, we can replace the spatial consistency verification (which is the bottleneck of existing methods) by a simple matching of individual keypoint without any significant deterioration of performances (as shown in the preliminary experiments). Therefore, the approach seems particularly suitable for scalable applications of CBVIR (e.g. image retrieval in huge databases).

Additionally, the analysis of implementation details suggests that both levels of image retrieval can be prospectively merged into a single process based on inverted indexing (similarly to [12]). Nevertheless, this issue is not discussed in this paper.

References

1. Csurka, G., Bray, C., Dance, C., Fan, L., Wilamowski, J.: Visual categorization with bags of keypoints. In: Proc. ECCV 2004, Workshop on Statistical Learning in Computer Vision, Prague (2004) 1–22
2. Chum, O., Perdoch, M., Matas, J.: Geometric min-hashing: Finding a (thick) needle in a haystack. In: Proc. IEEE Conf. CVPR 2009. (2009) 17–24
3. Jegou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. International Journal of Computer Vision **87** (2010) 316–336
4. Tolias, M., Jegou, H.: Visual query expansion with or without geometry: refining local descriptors by feature aggregation. Pattern Recognition **47** (2014) 3466–3476
5. Śluzek, A.: Visual categorization with bags of contextual descriptors improving credibility of keypoint matching. In: Proc. ICARCV 2014, Singapore (2014) (in print).
6. Zhao, W.L., Ngo, C.W., Tan, H.K., Wu, X.: Near-duplicate keyframe identification with interest point matching and pattern learning. IEEE Transactions on Multimedia **9** (2007) 1037–1048
7. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. International Journal of Computer Vision **60** (2004) 63–86
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
9. Arandjelovic, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: Proc. IEEE Conf. CVPR 2012. (2012) 2911–2918
10. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. PAMI **27** (2005) 1615–1630
11. Nistér, D., Stewénius, H.: Scalable recognition with a vocabulary tree. In: Proc. IEEE Conf. CVPR 2006. Volume 2. (2006) 2161–2168
12. Stewénius, H., Gunderson, S., Pilet, J.: Size matters: Exhaustive geometric verification for image retrieval. In: Proc. ECCV 2012. Volume II., Florence (2012) 674–687
13. Zhang, Y., Jia, Z., Chen, T.: Image retrieval with geometry-preserving visual phrases. In: Proc. IEEE Conf. CVPR 2011. (2011) 809–816
14. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proc. 9th IEEE Conf. ICCV 2003. Volume 2., Nice (2003) 1470–1477
15. Cha, S.H., Srihari, S.: On measuring the distance between histograms. Pattern Recognition **35** (2002) 1355–1370
16. Swain, M., Ballard, D.: Color indexing. International Journal of Computer Vision **7** (1991) 11–32
17. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proc. IEEE Conf. CVPR 2007. (2007) 1–8
18. Paradowski, M., Śluzek, A.: Local Keypoints and Global Affine Geometry: Triangles and Ellipses for Image Fragment Matching. In: Innovations in Intelligent Image Analysis. Volume SCI339. Springer-Verlag (2011) 195–224
19. Śluzek, A., Paradowski, M.: Visual similarity issues in face recognition. International Journal of Biometrics **4** (2012) 22–37