Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 167–173

ISBN 978-83-60810-51-4

Is Visual Similarity Sufficient for Semantic Object Recognition?

Andrzej Śluzek
Khalifa University
Dept. of Electrical and Computer Engineering, Abu Dhabi

Mariusz Paradowski
Wroclaw University of Technology
Institute of Informatics, Poland

***Abstract*—The paper discusses experiments (using exemplary classes of man-made objects) on *the-same-class* object detection based on the keypoint matching techniques. Two algorithms are used, i.e. building clusters of consistently similar and distributed keypoints, and matching individual points represented by novel descriptors incorporating semi-local geometry of images. It is shown that although detection of near-identically looking objects in random images can be performed reliably, the same is not possible for semantically defined classes of objects (even if we expect a certain level of visual and configurational uniformity within the class). The experiments conducted on PASCAL2007 dataset provide results which are not better than random selection. However, selected experimental results indicate that for certain classes of objects semantics may be significantly correlated with the visual and configurational consistencies.**

I. Introduction

KEYPOINT-based techniques (matching descriptors or visual words, *bag-of-word* approaches, etc.) are a popular tool in CBVIR (*content-based visual information retrieval*). They have been successfully applied in several problems of gradually increased complexity. Originally, keypoint detection (and the subsequent matching of keypoint descriptors) was proposed primarily for the retrieval of images depicting basically the same scenes under some photometric and geometric (e.g. viewpoint changes) distortions, and possibly with minor modifications of image contents, e.g. [1], [2], [3], [4]. Retrieval of objects similar to the object template is sometimes considered a generalization of the previous application. There, the reference image depicts a model view of the object of interest, and other images may possibly contain similar objects. This problem is often refereed to as *sub-image retrieval*, e.g. [5], [6], [7].

The most difficult task is obviously *near-duplicate-fragment* detection where the objective is to retrieve images containing unspecified numbers of nearly identical (i.e. generally looking "the same" but possibly distorted geometrically and/or photometrically) fragments. Near-duplicate fragment retrieval usually combines keypoint matching with either local (e.g. [6], [8], [9]) or semi-local (e.g. [10], [11]) analysis of configuration constraints, which is needed to identify groups of similarly transformed keypoints.

Keypoints have also been used for more advanced image matching problems. It has been suggested that statistical properties of keypoint descriptors (typically, the descriptors are approximated by a finite set of *visual words*) can characterize objects and/or scenes which not necessarily are similar in a purely visual sense but are similar semantically. For example, histograms of word distribution (i.e. *bags-of-words*, BoW) were used to retrieve images containing views of semantically defined classes of objects (e.g. 5 classes in [12] or 7 classes in [13]). Unfortunately, the number of classes should be predefined and, additionally, such systems need training before they are able to identify images containing *the-same-class* objects.

As a further generalization, BoW approach has been applied to the *scene categorization*, e.g. [14], and even for *action recognition*, e.g. [15] (where the keypoint detection is performed in a spatio-temporal domain). In such problems, the actual locations of keypoints are sometimes irrelevant (for scene categorization, in particular) so that keypoints are often densely sampled (rather than sparsely detected) over regular grids, e.g. [16], [17]. Obviously, the list of categories is predefined and extensive training (involving advanced machine learning techniques) is necessary.

In this paper, we attempt to link the above two topics, i.e. the retrieval based on purely visual similarities and the semantic-based retrieval. We are certainly aware how difficult is to close such a "semantic gap" (see [18]) so that the experiments have been conducted under several assumptions restricting the scope of the problem.

Thus, the problem is specified as follows:

1) The objective is to retrieve (from a collection of unknown images) images containing semantically similar (i.e. *the-same-class*) objects placed within random scenes.
2) The classes are not predefined, but we generally consider only classes of man-made objects or other objects for which certain visual and geometric uniformities should exist (e.g. *car*, *chair*, *camera*, *airliner*, *tank*, etc.). The classes of typical natural objects (e.g. *tree*, *bird*, *mountain*, *stone*, etc.) and classes of highly non-uniform man-made objects (e.g. *dress*, *toy*, *bottle label*, *book cover*,etc.) are excluded.
3) There is no learning process, and no positive or negative examples of objects are provided.

Based on our experiences, we hope to identify such *the-same-class* objects by detecting fragments (possible very small ones) for which the visual and geometric similarities (represented by keypoints and their configurations) exist. It should

be noted that retrieval of images sharing semantically similar objects is an important step towards automatic (or semi-automatic) annotation; this is one of the fundamental problems in handling large visual dataset.

In Section II of this paper we briefly overview two underlying mechanisms of image matching (discussed in details in previous papers). First (in Subsection II-A) we present a model with semi-local configuration constraints (see [11]) while the second model (described in Subsection II-B) is based on matching individual features only (see [19]). The experimental verification is provided in Section III. We compare results using two publicly available databases, i.e. VISIBLE and PASCAL2007. Unfortunately, the results of these experiments are rather negative though certain positive aspects are found as well. The results are interpreted and discussed in Section IV which also provides recommendations for the future researches in this area.

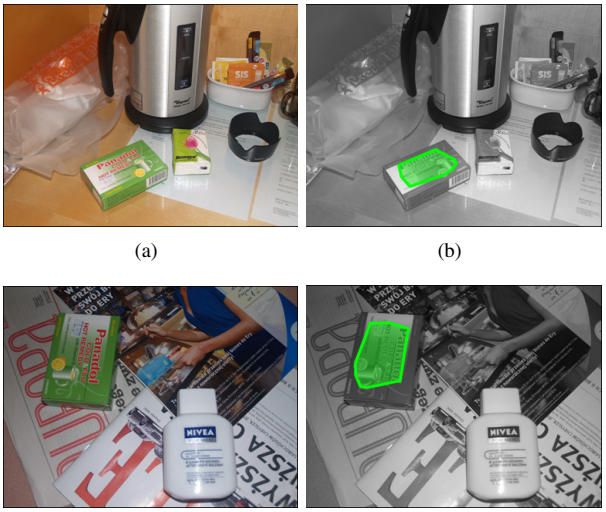
II. Models of Image Matching

A. Clusters of Similar Keypoints

Within the considered types of object classes, we assume some intra-class structural similarities (represented by configurations of correspondingly matched keypoints). Because very similar concepts (detection of clusters of similar and similarly distributed keypoints) was presented in [11], we use the same approach in this work. It is assumed in this approach that keypoints are first matched (using either *one-to-one* or *many-to-many* schemes) and, subsequently, affine transforms are randomly built between triplets (or pairs, if the keypoint shapes are exploited) of matching keypoints. Eventually, a histogram of affine parameters is constructed for the compared images. Clusters of keypoints related by similar transforms generate local maxima of the histogram so that, by detecting histogram spikes, we can both detect near-duplicate image fragments (even if they are distorted by affine mappings) and localize such fragments (using the coordinates of keypoints contributing to a cluster) within both images. Fig. 1 shows an exemplary pair of images with a near-duplicate fragment extracted by this method (details of the method are available in [11]).

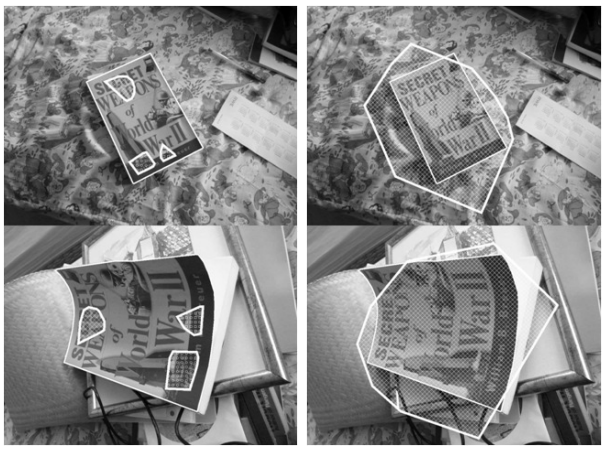
We have also implemented a modified variant where instead of geometrically consistent clusters of keypoints (represented by the histogram maxima) near-duplicates are defined by groups of topologically consistent (and correspondingly similar keypoints). This variant can tolerate more distortions of underlying objects so that it might be more suitable for classes of objects with more diversified configurations. An example illustrating typical differences between the original method and its topological variant is given in Fig. 2. Details of the topological variant are available in [20].

It is, therefore, assumed that views of *the-same-class* objects should contain such near-duplicates (either geometric or topological). Feasibility of this hypothesis is further investigated in Section III.



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

Fig. 1. A pair of images (a, c) and near-duplicate fragments (b, d) extracted using the histogram of affine transforms.



(a) (b)

Fig. 2. A pair of images matched using the histogram of affine transforms (a) and using the topological variant of the method (b).

B. TERM Features

As an alternative, images can be matched using novel *TERM* features which are discussed in [19]. In general, *TERM* features are combinations of quadrilaterals determined by the geometry of multi-ellipse configurations. As the most typical example, The principles of *TERM*3 (which are the most typical example of *TERM* features)are illustrated and explained in Fig. 3. First, it is shown how a trapezoid can be uniquely defined by an ellipse and two external points. Then, by locating other ellipses around these two external points, three trapezoids form a *TERM*3 feature. Such configurations of trapezoids change under affine transformations covariantly with the underlying ellipses. Thus, any affine-invariant shape descriptors computed over *TERM*3 features (we actually pro-

pose a simple descriptor based on moment invariants, see [19]) can be used to identify triplets of similarly distorted elliptic keypoints (note that the visual content of the ellipses are not analyzed).

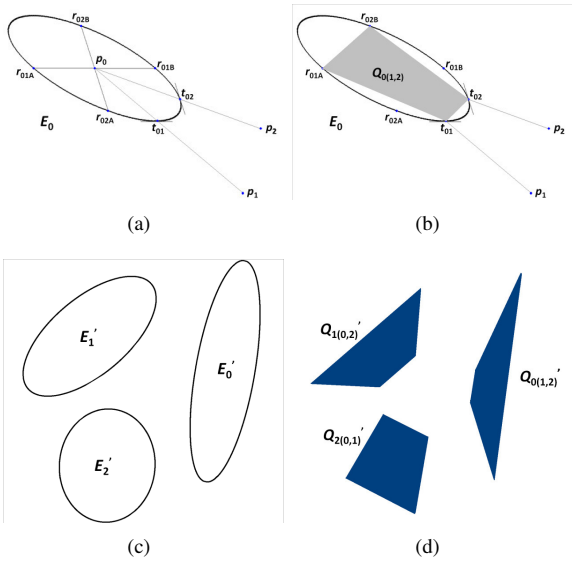


Fig. 3. Building a trapezoid in an ellipse in the context of two other points (a, b) and a configuration of three such trapezoids forming a *TERM*3 feature for three ellipses (c, d)).

In [19], keypoint configurations are jointly characterized both by words created from *TERM*3 descriptors (geometry) and by *SIFT* words (visual content) so that both the local visual characteristics and the semi-local image geometry are affine-invariantly represented. It has been shown that image matching based on correspondences between so described triplets of keypoints is more flexible; in particular more visual and geometric distortions can be tolerated. Therefore, this approach could be prospectively more appropriate (than the method presented in Subsection II-A) for retrieving images containing *the-same-class* objects.

Exemplary keypoint matching results (for images containing near-duplicate fragments) based on *TERM* words combined with *SIFT* words are shown in Fig. 4. More details of this method are available in [19].

III. Experiments

A. Methodology

The direct objective of the conducted experiments was to retrieve (within the available datasets) pairs of images containing fragments which are matched by the techniques briefly presented in Section II. It was believed that such matches would, indirectly, identify images containing *the-same-class* objects where the classes of objects are semantically defined (at least for selected classes of man-made objects for which a certain level of visual and configurational uniformity should exist). In contrast to already existing methods based on BoW approaches (and incorporating the learning phase for predefined types of objects, e.g. [12], [13]) the classes of objects

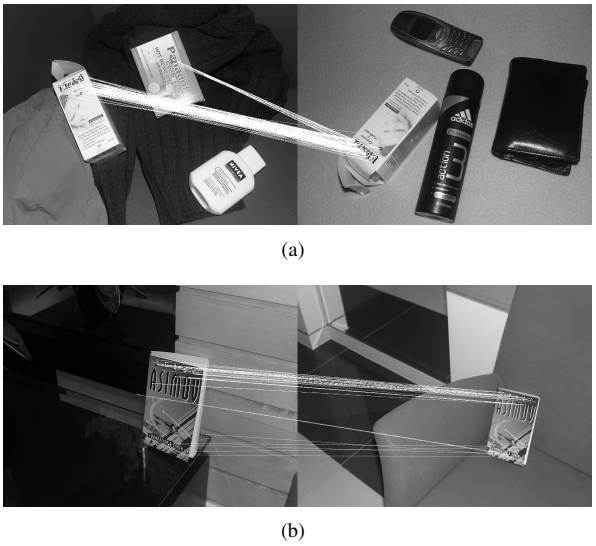


Fig. 4. Two examples of keypoint matching using a combination of *TERM* and *SIFT* words.

are not predefined and the image contents are unpredictable. Therefore, the statistics of visual word distributions cannot be learned for the classes existing in the available visual datasets. In other words, we attempt to retrieve images containing *the-same-class* objects (and to localize these objects within images) without any knowledge about the number of classes and their visual characteristics. The matching process is based only on visual characteristics and semi-local geometry.

Even if the above expectations are not fully confirmed, we at least hope to identify some classes of objects for which this approach works.

Two publicly available visual datasets (i.e. VISIBLE[1] and PASCAL2007[2]) are used. Note that we use only a part of PASCAL2007 (images containing selected man-main objects, i.e. *car*, *monitor*, *bus*, *plane*, *steam engine* and *train*). Nevertheless, the assumption about unpredictability of the image contents is not violated because the retrieval process does not utilize any preexisting knowledge about the number or types of classes.

The advantage of those two datasets is that they provide the ground truth in a form of manually outlined near-duplicate fragments (VISIBLE) or manually outlined *the-same-class* objects (PASCAL2007). Therefore, performances of image retrieval can be evaluated by comparing the results to the ground truth. In case of the histogram-based method, we count the relative number of detected near-duplicate fragments overlapping the ground-truth near-duplicate areas. In the second method (based on *TERM*3 features), the relative numbers of keypoint matches within the ground-truth outlines are counted. Exemplary images (and the corresponding ground truth) of both datasets are given in Fig. 5

[1]http://www.ii.pwr.wroc.pl/~visible/data/upload/FragmentMatchingDB.zip
[2]http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/VOCtrainval06-Nov-2007.tar

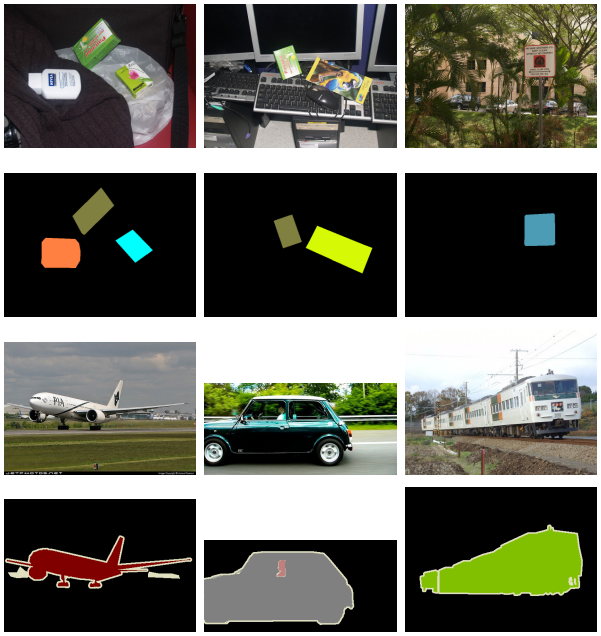


Fig. 5. Exemplary images of VISIBLE dataset (top rows) and PASCAL2007 dataset (bottom rows). The ground truth masks are also provided.

Within VISIBLE dataset, we can define *the-same-class* objects as identical copies of actually the same object. In other words, the class definition does not incorporate any semantics, and only purely visual characteristics are used. Therefore, the results obtained for VISIBLE can be considered a benchmark, and we should not expect better performances for classes defined by combinations of semantics and visual characteristics.

In the conducted experiments, we use Harris-Affine keypoint detectors (see [4]) and SIFT descriptors (which are popular standards). However, any affine-invariant keypoint detectors and any relevant descriptors can be alternatively employed.

B. Experiments using Clusters of Keypoints

Extensive experimental verification of the keypoint clusterization approach on VISIBLE database is available in [11]. The results (which have been also confirmed on other databases) are summarized in Table I.

TABLE I
PERFORMANCES OF NEAR-DUPLICATE FRAGMENT RETRIEVAL BASED ON CLUSTERING SIMILAR KEYPOINTS FOR VISIBLE DATASET.

Measure	Affine histograms	Topology
Precision	97%	97%
Recall	81%	94%

The content of the table confirms that almost all retrieved near-duplicate fragment actually indicate identically looking objects. However, not all pairs of identically looking objects are retrieved, especially by the affine histograms ($recall = 81\%$). This is because some objects are non-linearily distorted (so that affine mappings cannot model the shape changes) or there are just too few keypoints extracted within such objects. The first problem is partially rectified in the topologigal method (its *recall* is 94%).

Exemplary results (some have been already shown in Figs 1 and 2) are given in Fig. 6.

Fig. 6. Exemplary near-duplicate matches for VISIBLE dataset by affine histograms (top row) or topology (bottom row).

Unfortunately, the same experiments conducted on PASCAL2007 database are diametrically different. As shown in Table II, retrieval of images containing semantically similar objects (even if some level of visual and geometric similarity between such objects is expected) using clusters of similar and consistently configured (geometrically of topologically) keypoints is virtually impossible. A few examples of acceptable retrievals (from a very limited number of correctly retrieved pairs of objects) are given in Fig. 7.

TABLE II
PERFORMANCES OF NEAR-DUPLICATE FRAGMENT RETRIEVAL BASED ON CLUSTERING SIMILAR KEYPOINTS FOR A SUBSET OF PASCAL2007 IMAGES.

Measure	Affine histograms	Topology
Precision	5.2%	11%
Recall	0.7%	0.9%

C. Experiments using TERM Features

As an alternative, we have tested the approach based on matching individual features only. However, descriptions of the keypoint-based features incorporate both *TERM* words and *SIFT* words. Thus, certain information about semi-local configurations of images is available. Nevertheless, clusters of

(a) (b) (c)

Fig. 7. Acceptable examples of semantically similar objects identified in PASCAL2007 dataset by affine histograms (a, b) or topology (c).

corresponding keypoints of consistent (geometrical or topological) configurations are not built. This approach is, therefore, much faster (only individual features are matched) and more configurational and visual diversity can be prospectively tolerated within the views of *the-same-class* objects.

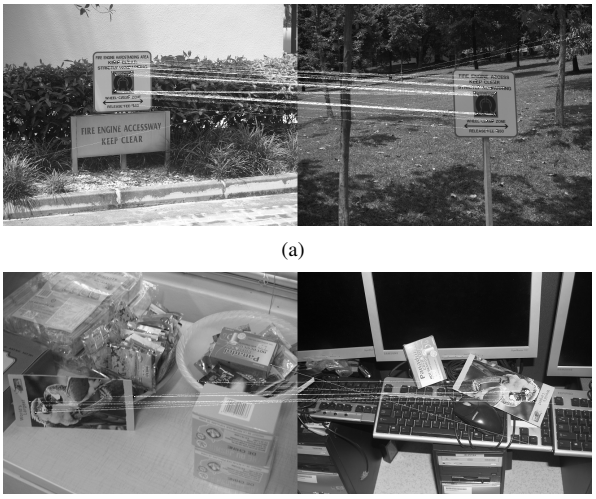
The measures of performance used in this experiment are slightly different than in the previous one. First, because we only match keypoint-based individual features, *precision* of keypoint correspondence is measured. We use an approximation where a match between two keypoints is considered true if in both images the keypoints belong to *the-same-class* ground truth outlines. The value of *recall* cannot be estimated because in a large dataset it is almost impossible to manually identify all ground-truth correspondences between keypoints. Secondly, we measure how reliably pairs of images are retrieved; a correct pair should contain *the-same-class* objects in both images. We use both *precision* and *recall*, but their values can fluctuate because we accept only pairs of images with the number of matched features exceeding a predefined threshold. If the threshold is low, *recall* might be higher, but *precision* deteriorates; if the threshold is high, the opposite happens. Thus, we try to maximize so-called F_β-*measure* (with a small value of β which highlights a lower importance of *recall*) and then use the corresponding values of *precision* and *recall*.

For the VISIBLE dataset, the results are relatively good (see Table III). Lower (compared with the clustering approach) performances of image retrieval are expected because individual feature matches (with more relaxed similarity criteria) may often indicate small similar patches, which do not belong to the actual objects.

TABLE III
PERFORMANCES OF *the-same-class* OBJECT RETRIEVAL USING KEYPOINT MATCHING BASED ON *TERM*3 AND *SIFT* WORDS (VISIBLE DATASET).

Measure	Precision	Recall	F_β-*measure*
Keypoint matching	91%	n/a	n/a
Pairs of images	87.5%	61%	85%

Exemplary matches found in VISIBLE dataset are provided in Fig. 8 (other examples have been already shown in Fig. 4). The examples illustrate why it would be difficult to obtain simultaneously high *precision* and *recall* in image pair retrieval. Matches (even if generally correct in case of images containing the same objects) often provide a small number of correspondences which represent random small patches which are only accidentally similar. Thus, a higher number of feature correspondences required to accept a pair of images is recommended to reject such random patch similarities. However, certain genuine matches (e.g. Fig. 8b) are also represented by relatively few correspondences. If a higher acceptance threshold is used, such pairs of images might be rejected.



(a)

(b)

Fig. 8. Exemplary keypoint matches in VISIBLE dataset obtained by combining *TERM*3 and *SIFT* words.

For PASCAL2007 dataset, however, the results are again (similarly to the first algorithm based of keypoint clustering) very disappointing. By randomly selecting *the-same-SIFT-word* keypoints (i.e. without using any structural description provided by *TERM*3 data) *precision* of keypoint matching is approx. 3%. Almost exactly the same value is obtained when the *TERM*3 data are used. In image pair retrieval, the highest *precision* is obtained by a random selection of image pairs (the value is approx. 15% and it is determined by the ratio between the number of pairs containing *the-same-class* objects and the total number of image pairs). The typical randomness of keypoint correspondences in images actually containing *the-same-class* objects is illustrated by selected examples given in Fig. 9.

IV. CONCLUSIONS

Superficially, the conducted experiments have been a failure. They clearly indicate that, in general, detection method for semantically defined *the-same-class* object is not feasible (even for classes with relatively uniform visual characteristics of objects) by using only visual properties of the objects not supported by any training process. This is also indirectly

Fig. 9. Exemplary keypoint matches in PASCAL2007 dataset obtained by combining TERM3 and SIFT word. Note that each pair of images contains *the-same-class* object.

confirmed in many other works where keypoints and/or visual words are used to identify images of *the-same-class* objects. All these works (e.g. [12], [13], [16], [21]) use classifiers built from collections of training data. In other words, systems learn the visual approximations of classes from a union of exemplary visual appearances (and possibly use the union of negative examples as well). Subsequently, input data are assigned to the class which contains appearances which are most similar to the visual query. If classes of objects are not predefined and their visual characteristics are not learnt (i.e. the system is not trained to recognize the visual diversity of *the-same-class* objects) the intra-class and the inter-class visual similarity/differences might be indistinguishable.

Nevertheless, it was shown in our previous work [22] how to automatically build visual classes from multiple near-duplicate fragments. Moreover, in spite of rather unsuccessful experiments on PASCAL2007 dataset, numerous examples of semantic similarities coexisting with large numbers of generally correct keypoint correspondences have been found using the *TERM*-based approach (even though the actual visual near-duplicity between the objects does not exist). Some of these examples are shown in Fig. 10.

We believe, therefore, that the investigations should be continued. Apparently, there are some classes of objects for which the semantic similarities are correlated with visual similarities significantly enough to automatically identify/define such classes of objects by visual analysis only, i.e. without any prior knowledge about the objects and their semantic classification.

Fig. 10. Examples of semantically similar objects with significant keypoint correspondences.

In contrast to [22], such automatically defined classes would not correspond to identically looking objects. Instead, they could indicate objects with systematically appearing visually near-duplicate structures.

The alternative approach is also possible. Classes of objects defined by visual similarities only may help to better understand problems of automatic image annotation (i.e. how to select semantically defined classes of object). The proposed annotation tags should just more systematically combine semantics of objects with their visual properties. For example, instead of *train* class, it might be more accurate to use *suburban train*, *freight train*, *passenger train*, etc. classes. Such a process could prospectively converge and the optimally-defined classes of objects could be built. Ideally, for such classes the visual characteristics would be fully correlated with the semantics.

REFERENCES

[1] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans PAMI*, vol. 19, no. 5, pp. 530–535, 1997.

[2] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. British Machine Vision Conference*, Cardiff, 2002, pp. 384–393.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, pp. 63–86, 2004.

[5] N. Sebe, M. Lew, and D. Huijsmans, "Multi-scale sub-image search," in *Proc. of 7th ACM Int. Conf. on Multimedia*, Orlando, FL, 1999, pp. 79–82.

[6] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Computer Vision*, vol. 2, 1999, pp. 1150–1157.

[7] Y. Ke, R. Sukthankar, and L. Huston, "Efficient near-duplicate detection and sub-image retrieval," in *Proc. ACM Multimedia Conf.*, 2004, pp. 869–876.

[8] O. Chum and J. Matas, "Matching with prosac - progressive sample consensus," in *Proc. IEEE Conf. CVPR 2005*, San Diego(CA), 2005, pp. 220–226.

[9] O. Chum, M. Perdoch, and J. Matas, "Geometric min-hashing: Finding a (thick) needle in a haystack," in *Proc. IEEE Conf. CVPR 2009*, 2009, pp. 17–24.

[10] W.-L. Zhao and C.-W. Ngo, "Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection," *IEEE Trans. on Image Processing*, vol. 2, pp. 412–423, 2009.

[11] M. Paradowski and A. Śluzek, *Innovations in Intelligent Image Analysis*. Springer-Verlag, 2011, vol. SCI339, ch. Local Keypoints and Global Affine Geometry: Triangles and Ellipses for Image Fragment Matching, pp. 195–224.

[12] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. IEEE Conf. CVPR 2003*, Madison, WI, 2003, pp. 264–271.

[13] G. Csurka, C. Bray, C. Dance, L. Fan, and J. Willamowski, "Visual categorization with bags of keypoints," in *Proc. 8th ECCV 2004, Workshop on Statistical Learning in Computer Vision*, Prague, 2004, pp. 1–22.

[14] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. CVPR 2005*, San Diego, CA, 2005, pp. 524–531.

[15] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. 2nd Joint IEEE Int. Workshop VS-PETSV*, Beijing, 2005, pp. 65–72.

[16] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proc. 9th ECCV 2006*, vol. LNCS 3954, Graz, 2006, pp. 490–503.

[17] B. Khadem, E. Farahzadeh, D. Rajan, and A. Śluzek, "Embedding visual words into concept space for action and scene recognition," in *Proc. British Machine Vision Conference*, Aberystwyth, 2010, pp. 15.1–15.11.

[18] H. Kwasnicka, M. Paradowski, M. Stanek, M. Spytkowski1, and A. Śluzek, "Image similarities on the basis of visual content an attempt to bridge the semantic gap," in *Proc. 3rd Int. Conf. ACIIDS 2011*, vol. LNAI 6591, Daegu, 2011, pp. 14–26.

[19] A. Śluzek and M. Paradowski, "Detection of near-duplicate patches in random images using keypoint-based features," in *Proc. ACIVS 2012*, vol. LNCS(in print), Brno, 2012.

[20] M. Paradowski and A. Śluzek, "Keypoint-based detection of near-duplicate image fragments using image geometry and topology," in *Proc. ICCVG'2010*, vol. LNCS 6375(2), Warsaw, 2010, pp. 175–182.

[21] Y. Su and F. Jurie, "Improving image classification using semantic attributes," *International Journal of Computer Vision (preprint)*, 2012.

[22] M. Paradowski and A. Śluzek, "Automatic visual class formation using image fragment matching," in *Proc. 5th Int. Symp. AAIA'10*, Wisla, 2010, pp. 97–104.