Spatially-sensitive affine-invariant image descriptors

Alexander M. Bronstein[1] and Michael M. Bronstein[2]

[1] Dept. of Electrical Engineering, Tel Aviv University
[2] Dept. of Computer Science, Technion – Israel Institute of Technology

Abstract. Invariant image descriptors play an important role in many computer vision and pattern recognition problems such as image search and retrieval. A dominant paradigm today is that of "bags of features", a representation of images as distributions of primitive visual elements. The main disadvantage of this approach is the loss of spatial relations between features, which often carry important information about the image. In this paper, we show how to construct *spatially-sensitive* image descriptors in which both the features and their relation are affine-invariant. Our construction is based on a vocabulary of pairs of features coupled with a vocabulary of invariant spatial relations between the features. Experimental results show the advantage of our approach in image retrieval applications.

1 Introduction

Recent works [1–9] demonstrated that images can be efficiently represented and compared using local features, capturing the most distinctive and dominant structures in the image. The construction of a feature-based representation of an image typically consists of *feature detection* and *feature description*, often combined into a single algorithm. The main goal of a feature detector is to find stable points or regions in an image that carry significant information on one hand and can be repeatedly found under transformations. Transformations typically considered include scale [3, 4], rotation, and affine [7, 8] transformations. A feature descriptor is constructed using local image information in the neighborhood of the feature points (or regions).

One of the advantages of feature-based representations is that they allow to think of images as a collection of primitive elements (*visual words*), and hence appeal to the analogy of text search and use well-developed methods from that community. Images can be represented as a collection of visual words indexed in a "visual vocabulary" by vector quantization in the descriptor space [10, 11]. Counting the frequency of the visual word occurrence in the image, a representation referred to as a *bag of features* (analogous to a *bag of words* used in search engines) is constructed. Images containing similar visual information tend to have similar features, and thus comparing bags of features allows to retrieve similar images.

Using invariant feature detectors and descriptors, invariance is built into bags of features by construction. For example, given two images differing by an affine transformation, their bag of features representations based on MSER descriptors are (at least theoretically) equal. Yet, one of the main disadvantages of bags of features is the fact that they consider only the statistics of visual words and lose the spatial relations between them. This may often result in loss of discriminativity, as spatial configuration of features often carries important information about the underlying image [12]. A similar problem is also encountered in text search problems. For example, in a document about "matrix decomposition" the word "matrix" is frequent. Yet, a document about the movie *Matrix* will also contain this word, which will result in a similar word statistics and, consequently, similar bags of features. In the most pathological case, a random permutation of words in a text will produce identical bags of words. In order to overcome this problem, text search engines commonly use vocabularies consisting not only of single words but also of combinations of words or *expressions*.

This text analogy can be extended to images. Unlike text which is one-dimensional, *visual expressions* are more complicated since the spatial relations of objects in images are two-dimensional. A few recent papers tried to extend bags of features taking into consideration spatial information about the features. Marszalek and Schmid [13] used spatial weighting to reduce the influence of background clutter (a similar approach was proposed in [14]). Grauman and Darrell [15] proposed comparing distributions of local features using *earth mover's distance* (EMD) [16], which incorporates spatial distances. Nister and Stewenius [17] used feature grouping to increase the discriminativity of image descriptors, and also showed that such the advantage of such an approach over enlarging the descriptor area is smaller sensitivity to occlusion. A similar approach for feature grouping and geometry consistency verification has been more recently proposed by Wu *et al.* [18]. Sivic *et al.* [19, 20] used feature configurations for object retrieval. Chum and Matas [21] considered a special case when the feature appearance is ignored and only geometry of feature pairs is considered. In [22], the spatial structure of features was captured using a multiscale bag of features construction. The representation proposed in [23] used spatial relations between parts. In [24], in a different application of 3D shape description, spatially-sensitive bags of features based on pairs of words were introduced. Behmo *et al.* [25] proposed a *commute graph* representation partially preserving the spatial information. However, the commute graph based on Euclidean distance relations is not invariant under affine transformations. Moreover, commute graphs encode only translational relations between features, ignoring more complicated relations such as scale and orientation of one feature with respect to another.

The main focus of this paper is the construction of affine-invariant feature-based image descriptors that incorporate spatial relations between features. Our construction is based on a vocabulary of pairs of features coupled with a vocabulary of affine-invariant spatial relations. Such a construction is a meta-approach which can augment existing feature description methods and can be considered as

an extension of the classical bags of features. The rest of the paper is organized as follows. In Section 2, we introduce notation and the notions of invariance and covariance, using which we formally define feature detection, description, and bags of features. Section 3 describes our construction of affine-invariant spatially-sensitive bags of features. Section 4 demonstrates the performance of our approach in an invariant image retrieval experiment. Finally, Section 5 concludes the paper.

2 Background

Typically, in the computation of a bag of features representation of an image, first a *feature detector* finds stable regions in the image. Next, each of the detected features undergoes is transformed to an invariant *canonical representation*, from which a *visual descriptor* is computed. Each such descriptor containing visual information about the feature is quantized in a *visual vocabulary*, increasing the count of the visual word corresponding to it. Finally, counts from all features are collected into a single distribution, called a *bag of features*. In what follows, we formalize each of these steps.

Feature detection. Let us be given an image I (for simplicity, grayscale). We refer to a planar subset F as to a *feature*, and denote by $\mathbf{F}_I = \{F_1, \dots, F_n\}$ a *feature transform* of I that produces a collection of features out of an image. The feature transform is said to be *covariant* with a certain group of geometric transformations if it commutes with action of the group, i.e., for every transformation $\mathbf{T}$, $\mathbf{F}_{\mathbf{T}I} = \mathbf{T}\mathbf{F}_I$ (we write $\mathbf{T}I(x)$ implying $I(\mathbf{T}x)$). In particular, we are interested in the group of affine transformations of the plane. We will henceforth assume that the feature transform is affine-covariant. A popular example of such a feature transform is MSER [7], which will be adopted in this study.

Feature canonization. Once features are detected, they are often normalized or *canonized* by means of a transformation into some common system of coordinates [26]. We denote the inverse of such a canonizing transformation associated with a feature F by $\mathbf{A}_F$, and refer to $\mathbf{A}_F^{-1}F$ as to a *canonical representation* of the feature. As before, this process is said to be affine-covariant if it commutes with the action of the affine group. The canonical representation in that case is *affine-invariant*, i.e., $\mathbf{A}_F^{-1}F = \mathbf{A}_{\mathbf{T}F}^{-1}(\mathbf{T}F)$ for every affine transformation $\mathbf{T}$. A classical affine-covariant (up to reflection ambiguity) feature canonization is based on zeroing its first-order moments (centroid) and diagonalizing the second-order moments [27].

Feature descriptors. The fact that a canonical representation of a feature is invariant is frequently used to create invariant descriptors. We will denote by $\mathbf{v}_F$ a vector representing the visual properties of the image supported on F and transformed by $\mathbf{A}_F^{-1}$ into the canonical system of coordinates, referring to it as to a *visual descriptor* of F. A straightforward descriptor can be obtained by simply sampling the feature footprint in the canonical space and representing the obtained samples in a vector form [26]. However, because of using the intensity values of the image directly, such a descriptor is sensitive to changes

in illumination. While this is not an issue in some applications, many real applications require more sophisticated representations. For example, the SIFT descriptor [3] computes a histogram of local oriented gradients (8 orientation bins for each of the 4×4 location bins) around the interest point, resulting in a 128-dimensional vector. SURF [9] descriptor is similar to SIFT yet more compact, with 4-dimensional representation for each of the 4×4 spatial locations (total of 64 dimensions).

Bags of features. Given an image, descriptors of its features are aggregated into a single statistic that describes the entire image. For that purpose, descriptors are vector-quantized in a *visual vocabulary* $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ containing m representative descriptors, which are usually found using clustering algorithms. We denote by $\mathbf{Q}_\mathbf{V}$ a quantization operator associated with the visual vocabulary $\mathbf{V}$ that maps a descriptor into a distribution over $\mathbf{V}$, represented as an m-dimensional vector. The simplest hard quantization is given by

$$(\mathbf{Q}_\mathbf{V}\mathbf{v})_i = \begin{cases} 1 : d(\mathbf{v}, \mathbf{v}_i) \leq d(\mathbf{v}, \mathbf{v}_j), \quad j = 1, \ldots, m \\ 0 : \text{else}, \end{cases} \tag{1}$$

where $d(\mathbf{v}, \mathbf{v}')$ is the distance in the visual descriptor space, usually the Euclidean distance $\|\mathbf{v} - \mathbf{v}'\|$. Summing the distributions of all features,

$$\mathbf{B}_I = \sum_{F \in \mathbf{F}_I} \mathbf{Q}_\mathbf{V}\mathbf{v}_F,$$

yields an affine-invariant representation of the image called a *bag of features*, which with proper normalization is a distribution of the image features over the visual vocabulary. Bags of features are often L_2-normalized and compared using the standard Euclidean distance or correlation, which allows efficient indexing and comparison using search trees or hash tables [11].

3 Spatially-sensitive image descriptors

A major disadvantage of bags of features is the fact that they discard information about the spatial relations between features in an image. We are interested in *spatially-sensitive* bags of features that encode spatial information in an invariant manner. As already mentioned in the introduction, spatial information in the form of expressions is useful in disambiguating different uses of a word in text search. A 2D analogy of two text documents containing the same words up to some permutation is a scene depicting different arrangements or motion of the same objects: a change in the relative positions of the objects creates different spatial configuration of the corresponding features in the image. Yet, in images, the spatial relations can also change as a result of a difference in the view point (usually approximated by an affine transformation). If in the former case the difference in spatial relations is desired since it allows us to discriminate between different visual content, in the latter case, the difference is undesired since it would deem distinct a pair of visually similar images.

Visual expressions. A straightforward generalization of the notion of combinations of words and expressions to images can be obtained by considering *pairs* of features. For this purpose, we define a visual vocabulary on the space of pairs of visual descriptors as the product $\mathbb{V} \times \mathbf{V}$, and use the quantization operator $\mathbf{Q}^2_{\mathbf{V}} = \mathbf{Q}_{\mathbf{V}} \times \mathbf{Q}_{\mathbf{V}}$ assigning to a pair of descriptors a distribution over $\mathbf{V} \times \mathbf{V}$. $(\mathbf{Q}^2_{\mathbf{V}}(\mathbf{v}, \mathbf{v}'))_{ij}$ can be interpreted as the joint probability of the pair $(\mathbf{v}, \mathbf{v}')$ being represented by the expression $(\mathbf{v}_i, \mathbf{v}_j)$.

Same way as expressions in text are pairs of adjacent words, visual expressions are pairs of spatially-close visual words. The notion of proximity can be expressed using the idea of canonical neighborhoods: fixing a disk M of radius $r > 0$ centered at the origin of the canonical system of coordinates, we define $N_F = \mathbf{A}_F M$ to be a *canonical neighborhood* of a feature F. Such a neighborhood is affine-covariant, i.e., $N_{\mathbf{T}F} = \mathbf{T}N_F$ for every affine transformation $\mathbf{T}$. The notion of a canonical neighborhood induces a division of pairs of features into near and far. We define a *bag of pairs of features* simply as the distribution of near pairs of features,

$$\mathbf{B}^2_I = \sum_{F \in \mathbf{F}_I} \sum_{F' \in N_F} \mathbf{Q}^2_{\mathbf{V}}(\mathbf{v}_F, \mathbf{v}_{F'}).$$

Bags of pairs of features are affine-invariant by their construction, provided that the feature transform and the canonization are affine-covariant.

Spatial relations. Canonical neighborhoods express binary affine-invariant proximity between features, which is a simple form of spatial relations. A more general class of spatial relations can be obtained by considering the relation between the canonical transformations of pairs of features. Specifically, we consider the *canonical relation*

$$\mathbf{S}_{F,F'} = \mathbf{A}^{-1}_{F'} \mathbf{A}_F.$$

It is easy to show that $\mathbf{S}_{F,F'}$ is affine-invariant, i.e., $\mathbf{S}_{\mathbf{T}F,\mathbf{T}F'} = \mathbf{S}_{F,F'}$ for every affine transformation $\mathbf{T}$. This spatial relation can be thought of as the transformation from F' to F expressed in the canonical system of coordinates. It should not be confused with the transformation from the system of coordinates of F' to the system of coordinates of F, which is achieved by $\mathbf{A}_F \mathbf{A}^{-1}_{F'}$.

It is worthwhile noting that symmetric features result in ambiguous spatial relations. The problem can be resolved by projecting the relation onto the subgroup of the affine group modulo the ambiguity group. When the ambiguity group is finite (e.g. reflection), the spatial relation can be defined as a set [28].

Spatially-sensitive bags of features. Being an invariant quantity, the canonical spatial relation can be used to augment the information contained in visual descriptors in a bag of pairs of features. For that purpose, we construct a vocabulary of spatial relations, $\mathbf{S} = \{\mathbf{S}_1, \ldots, \mathbf{S}_n\}$. A quantization operator $\mathbf{Q}_{\mathbf{S}}$ associated with the spatial vocabulary can be constructed by plugging an appropriate metric into (1). The easiest way of defining a distance on the space of transformations is the Frobenius norm on transformations represented in homo-

geneous coordinates,

$$d^2(\mathbf{S},\mathbf{S}') = \|\mathbf{S}-\mathbf{S}'\|_{\mathrm{F}}^2 \;=\; \mathrm{tr}((\mathbf{S}-\mathbf{S}')^{\mathrm{T}}(\mathbf{S}-\mathbf{S}')),$$

which is equivalent to considering the 3×3 transformation matrices as vectors in $\mathbb{R}^9$ using the standard Euclidean distance. A somewhat better approach is to use the intrinsic (geodesic) distance on the Lie group of matrices,

$$d^2(\mathbf{S},\mathbf{S}') = \|\log(\mathbf{S}^{-1}\mathbf{S}')\|_{\mathrm{F}}^2,$$

where $\log\mathbf{X} = \sum_{i=0}^{\infty}\frac{(-1)^{i+1}}{i}(\mathbf{X}-\mathbf{I})^i$ is the matrix logarithm.

The disadvantage of the intrinsic distance is the non-linearity introduced by the logarithm. However, using the Baker-Campbell-Hausdorff exponential identity for non-commutative Lie groups yields the following first-order approximation,

$$\begin{aligned} d(\mathbf{S},\mathbf{S}') = \|\log(\mathbf{S}^{-1}\mathbf{S}')\|_{\mathrm{F}} &= \big\|\log\big(\exp(-\log\mathbf{S})\exp(\log\mathbf{S}')\big)\big\|_{\mathrm{F}} \\ &= \big\|\log\big(\exp(\log\mathbf{S}') - \exp(\log\mathbf{S}) + \mathcal{O}(\|\log\mathbf{S}'\log\mathbf{S}\|^2)\big)\big\|_{\mathrm{F}} \\ &\approx \|\log\mathbf{S}' - \log\mathbf{S}\|_{\mathrm{F}}. \end{aligned}$$

Practically, using this approximation, spatial relations can be thought of as nine-dimensional vector whose elements are the entries of the logarithm matrix $\log\mathbf{S}$, and the distance between them is the standard Euclidean distance on $\mathbb{R}^9$. A more general distance between spatial relations can be obtained by projecting $\mathbf{S}$ and $\mathbf{S}'$ onto subgroups of the affine group, measuring the distances between projections, and then combining them into a single distance.

Coupling the spatial vocabulary $\mathbf{S}$ with the visual vocabulary $\mathbf{V}\times\mathbf{V}$ of pairs of features, we define the *spatially-sensitive bag of features*

$$\mathbf{B}_I^3 = \sum_{F\in\mathbf{F}_I}\sum_{F'\in N_F}\mathbf{Q}_{\mathbf{V}}^2(\mathbf{v}_F,\mathbf{v}_{F'})\cdot\mathbf{Q}_{\mathbf{S}}(\mathbf{S}_{F,F'}),$$

which, with proper normalization, is a distribution over $\mathbf{V}\times\mathbf{V}\times\mathbf{S}$ that can be represented as a three-dimensional matrix of size $m\times m\times n$. Spatially-sensitive bags of features are again affine-invariant by construction.

While the clear advantage of spatially-sensitive bags of features is their higher discriminativity, the resulting representation size may be significantly higher. Additional potential drawback is that repeatability of pairs of features can be lower compared to single features. Due to the above considerations, the best application for the presented approach is a scenario in which the two images to be compared have a large overlap in the visual content. An example of such an application is image and video copy detection, in which one tries to recognize an image or video frame that has undergone some processing or tampering. Another example is video alignment, in which one tries to find a correspondence between two video sequences based on their visual content. Subsequent frames in video may differ as a result from motion, which result in different spatial configurations of the depicted object. Distinguishing between such frames using bags of features would be very challenging or even impossible (see e.g. Figure 2).

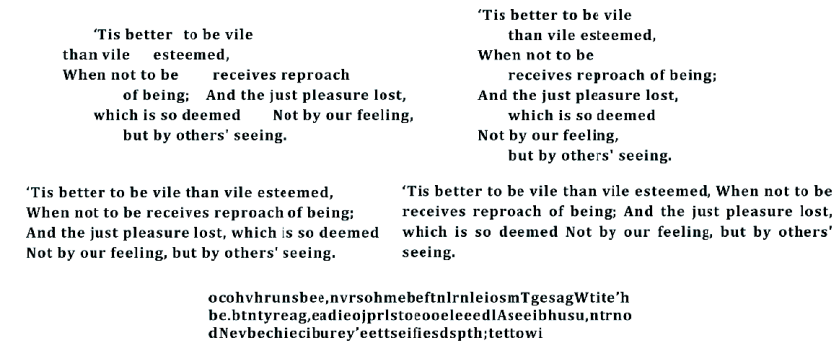

Fig. 1. Examples of five layouts of a Shakespearean sonnet from the *Text* dataset. The last layout is a random permutation of letters.

Fig. 2. Examples of three images from the same scene in the *Opera* dataset. Each scene contains visually similar objects appearing in different spatial configurations. Such images are almost indistinguishable by means of bags of features, yet, result in different spatially-sensitive descriptors.

4 Results

We assessed the proposed methods in three image retrieval experiment, using *Text*, *Opera*, and *Still life* datasets described in the following.[3] The first two experiments were with synthetic transformations, the third experiment was with real photographed data. The datasets were created to contain objects in different geometric configurations. In all the experiments, MSER was used as the feature detector, followed by the moment-based canonization. Feature descriptors were created by sampling the unit square in the canonical space on a 12×12 grid. Three methods were compared: simple bags of features (BoF), bags of pairs of features (P-BoF), and spatially-sensitive bags of features (SS-BoF). All bags of features were computed from the same sets of feature descriptors and canonical transformations using the same visual vocabularies.

Synthetic data. The first two experiments were performed on two datasets. The first was the *Text* dataset consisting of 29 distinct fragments from Shake-

[3] All the data and code for reproducing the experiments will be published online.

Fig. 3. Examples of three viewpoints (left, middle, right) and two configurations (first and second rows) of objects in the Still life dataset. Images in the same layout were taken by multiple cameras from different positions.

spearian sonnets. Each fragment was rendered as a black-and-white image using the same font in several spatial layouts containing the same letters organized differently in space. One of such extreme layouts included a random permutation of the letters. This resulted in a total of 91 images, a few examples of which are depicted in Fig. 1. Black-and-white text images are an almost ideal setting for the MSER descriptor, which manifested nearly perfect affine-invariance. This allowed to study in an isolated manner the contribution of spatial relations to bag of feature discriminativity.

The *Opera* dataset was composed of 28 scenes from different opera recordings. From each scene, several frames were selected in such a way to include approximately the same objects in different spatial configurations, resulting in a total of 83 images (Fig. 2). The challenge of this data was to be able to distinguish between different spatial configurations of the objects. Such a problem arises, for example, in video alignment where subsequent frames are often very similar visually but have slightly different spatial layouts.

To each image in both data sets, 21 synthetic transformation were applied. The transformations were divided into five classes: in-plane rotation, mixed in-plane and out-of-plane rotation, uniform scaling, non-uniform scaling, and null (no transformation). Each transformation except the null appeared in three increasing strengths (marked $1-5$).

For the *Text* data, the vocabularies were trained on examples of other text, not used in the tests. Same visual vocabulary of size 128 were used in all the algorithms; spatial vocabulary of size 24 was used in SS-BoFs. For the *Opera* data, the vocabularies were trained on web images. Visual vocabulary was of

Method	Transformation	Strength 1	≤2	≤3	≤4	≤5
BoF	*In-plane rotation*	41.57	35.33	31.98	30.86	30.39
	Mixed rotation	26.28	35.23	32.68	28.56	24.25
	Nonuniform scale	58.13	59.70	59.25	60.11	58.63
	Uniform scale	55.30	48.64	46.79	45.77	44.57
	All	45.32	44.73	42.68	41.33	39.46
P-BoF	*In-plane rotation*	60.51	49.36	43.45	40.90	39.94
	Mixed rotation	30.08	48.86	42.97	36.05	30.33
	Nonuniform scale	81.90	82.90	83.13	83.26	81.31
	Uniform scale	78.91	72.56	73.06	69.82	67.73
	All	62.85	63.42	60.65	57.51	54.83
SS-BoF	*In-plane rotation*	100.00	100.00	99.45	99.08	99.12
	Mixed rotation	97.99	98.99	98.14	85.96	70.48
	Nonuniform scale	100.00	100.00	100.00	100.00	100.00
	Uniform scale	100.00	100.00	100.00	100.00	100.00
	All	99.50	99.75	99.40	96.26	92.40

Table 1. Retrieval performance (mAP in %) of different methods on the *Text* dataset, broken down according to transformation classes and strengths (1–5).

size 128, and spatial vocabulary was of size 24. In all experiments, the size of the canonical neighborhood was set to $r = 15$.

We performed a leave-one-out retrieval experiment on both datasets. Euclidean distance between different image descriptors (BoF, P-BoF, and SS-BoF) was used to rank the results. Retrieval performance was evaluated on subsets of the distance matrix using precision/recall characteristic. *Precision at* k, $P(k)$, is defined as the percentage of relevant images in the first k top-ranked retrieved images. Relevant images were the same configuration of objects regardless of transformation. *Average precision* (AP) is defined as $mAP = \frac{1}{R}\sum_k P(k)\cdot rel(k)$, where $rel(k) \in \{0, 1\}$ is the relevance of a given rank and R is the total number of relevant images. *Mean average precision* (mAP), the average of AP over all queries, was used as a single measure of retrieval performance. Ideal retrieval results in all first matches relevant (mAP=100%).

Tables 1 and 2 shows the retrieval performance using different image representations on *Text* and *Opera* datasets, respectively. The performance is broken down according to transformation classes and strengths. The use of spatially-sensitive bags of features increases the performance from 39.46% mAP to 92.4% (134% improvement) on the *Text* data and from 83.9% to 91.35% (8% improvement) on the *Opera* data.

Real data. In the third experiment, we used the *Still life* dataset containing 191 images of objects laid out in 9 different configurations (scenes) and captured from multiple views with very wide baseline by cameras with different focus and resolution (12–36 views for each scene). Some of the views differed dramatically, including occlusions, scene clutter, as shown in Figure 3. Moreover, most of the scenes included a sub-set of the same objects. The challenge in this experiment

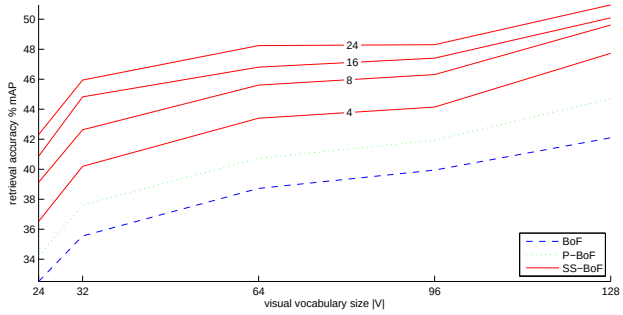

Fig. 4. Performance of different methods on the *Still life* dataset. Four red solid curves correspond to SS-BoF with spatial vocabulary of different size (displayed on the curve).

was to group images into scenes based on their visual similarity. Same vocabularies as in the *Opera* test were used. We performed a leave-one-out retrieval experiment. Successful match was from the same object configuration (e.g., in Figure 3, when querying the top left image, correct matches are top middle and right, incorrect matches are all images in the second row).

Figure 4 shows the retrieval accuracy of different methods as a function of visual and spatial vocabulary size. BoF achieves the best retrieval performance (42.1% mAP) with a vocabulary of size 128. With the same vocabulary, P-BoF achieves 44.7% mAP. The best result for SS-BoF is 51.0% (21% improvement) when using a spatial vocabulary of size 24. Consistent and nearly constant improvement is exhibited for all the range of the tested visual vocabulary sizes.

We observe that while consistent improvement is achieved on all datasets, spatially-sensitive bags of features perform the best on the *Text* data. We attribute this in part to the relatively primitive feature canonization method used in our experiments, which was based only on the feature shape and not on the feature intensity content. This might introduce noise into the computed canonical transformation and therefore degrade the performance of canonical neighbors and spatial vocabulary. In future studies, we intend to use a SIFT-like feature canonization based on the dominant intensity direction, which is likely to improve the stability of the canonical transformations.

5 Conclusions and future directions

We presented a construction of a feature-based image representation that generalizes the bag of features approach by taking into consideration spatial relations between features. Central to our construction is a vocabulary of pairs of affine-invariant features coupled with a vocabulary of affine-invariant spatial relations. The presented approach is a meta-algorithm, since it augments the standard bag of features approach and is not limited to a specific choice of a feature trans-

Method	Transformation	Strength 1	≤2	≤3	≤4	≤5
	In-plane rotation	92.95	88.36	84.62	80.63	77.59
	Mixed rotation	68.39	64.98	69.86	70.25	70.07
SS-BoF	*Nonuniform scale*	95.50	96.01	95.90	95.22	94.47
	Uniform scale	96.73	95.16	94.78	94.31	93.47
	All	88.39	86.13	86.29	85.10	83.90
	In-plane rotation	93.55	88.19	85.82	84.17	81.49
	Mixed rotation	75.42	72.84	75.47	75.21	74.75
SS-BoF	*Nonuniform scale*	95.31	96.01	95.53	95.13	94.52
	Uniform scale	96.18	94.76	93.86	93.62	93.09
	All	90.11	87.95	87.67	87.03	85.96
	In-plane rotation	95.11	92.73	91.27	89.91	88.52
	Mixed rotation	82.68	81.42	83.65	83.95	84.23
SS-BoF	*Nonuniform scale*	97.32	97.32	97.64	97.36	96.47
	Uniform scale	98.80	98.11	97.28	96.66	96.20
	All	93.48	92.40	92.46	91.97	91.35

Table 2. Retrieval performance (mAP in %) of different methods on the *Opera* dataset, broken down according to transformation classes and strengths (1–5).

form. In future studies, we intend to test it on other descriptors such as SIFT, and extend the idea of spatial relations to epipolar relations between features in calibrated images. We also intend to extend the proposed approach to video, creating affine-invariant vocabularies for motion.

Experimental results show improved performance of image retrieval on synthetic and real data. We plan to evaluate our approach in a large-scale image retrieval experiment. Our approach is especially suitable for problems in which the compared images have large overlap in visual content, such as copy detection and video alignment, an application that will be studied in future works.

References

1. Lindeberg, T.: Feature detection with automatic scale selection. IJCV **30** (1998) 79–116
2. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: Proc. ICCV. Volume 1. (2001) 525–531
3. Lowe, D.: Distinctive image features from scale-invariant keypoint. IJCV (2004)
4. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. IJCV **60** (2004) 63–86
5. Tuytelaars, T., Van Gool, L.: Matching widely separated views based on affine invariant regions. IJCV **59** (2004) 61–85
6. Kadir, T., Zisserman, A., Brady, M.: An affine invariant salient region detector. Lecture Notes in Computer Science (2004) 228–241
7. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing **22** (2004) 761–767

8. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.: A comparison of affine region detectors. IJCV **65** (2005) 43–72
9. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. Lecture Notes in Computer Science **3951** (2006) 404
10. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proc. CVPR. (2003)
11. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: Proc. ICCV. (2007)
12. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: Proc. CVPR. (2007) 1–8
13. Marszaek, M., Schmid, C.: Spatial weighting for bag-of-features. In: Proc. CVPR. Volume 2. (2006)
14. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: Workshop on Statistical Learning in Computer Vision, ECCV. (2004) 17–32
15. Grauman, K., Darrell, T.: Efficient image matching with distributions of local invariant features. In: Proc. CVPR. Volume 2. (2005)
16. Rubner, Y., Tomasi, C., Guibas, L.: The earth mover's distance as a metric for image retrieval. IJCV **40** (2000) 99–121
17. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Proc. CVPR. (2006)
18. Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling features for large-scale partial-duplicate web image search. In: Proc. CVPR. (2009)
19. Sivic, J., Zisserman, A.: Video data mining using configurations of viewpoint invariant regions. In: Proc. CVPR. (2004)
20. Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.: Discovering object categories in image collections. In: Proc. ICCV. Volume 2. (2005)
21. Chum, O., Matas, J.: Geometric hashing with local affine frames. In: Proc. CVPR. (2006)
22. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. CVPR. Volume 2. (2006)
23. Amores, J., Sebe, N., Radeva, P.: Context-based object-class recognition and retrieval by generalized correlograms. IEEE Trans. PAMI **29** (2007) 1818–1833
24. Ovsjanikov, M., Bronstein, A.M., Bronstein, M.M., Guibas, L.: Shape google: a computer vision approach to invariant shape retrieval. In: Proc. NORDIA. (2009)
25. Behmo, R., Paragios, N., Prinet, V.: Graph commute times for image representation. In: Proc. CVPR. (2008)
26. Forssén, P., Lowe, D.: Shape descriptors for maximally stable extremal regions. In: Proc. ICCV. (2007) 59–73
27. Muse, P., Sur, F., Cao, F., Lisani, J.L., Morel, J.M.: A theory of shape identification (2005)
28. Bronstein, A.M., Bronstein, M.M.: Affine-invariant spatial vocabularies. Technical Report Techn. Report CIS-2009-10, Dept. of Computer Science, Technion, Israel (2009)