Proximity Distribution Kernels for Geometric Context in Category Recognition

Haibin Ling*
Integrated Data Systems Department
Siemens Corporate Research, Princeton, NJ
haibin.ling @ siemens.com

Stefano Soatto
Computer Science Department
University of California, Los Angeles, CA
soatto @ cs.ucla.edu

Abstract

We propose using the proximity distribution of vector-quantized local feature descriptors for object and category recognition. To this end, we introduce a novel "proximity distribution kernel" that naturally combines local geometric as well as photometric information from images. It satisfies Mercer's condition and can therefore be readily combined with a support vector machine to perform visual categorization in a way that is insensitive to photometric and geometric variations, while retaining significant discriminative power. In particular, it improves on the results obtained both with geometrically unconstrained "bags of features" approaches, as well as with over-constrained "affine procrustes." Indeed, we test this approach on several challenging data sets, including Graz-01, Graz-02, and the PASCAL challenge. We registered the average performance at 91.5% on Graz-01, 82.7% on Graz-02, and 74.5% on PASCAL. Our approach is designed to enforce and exploit geometric consistency among objects in the same category; therefore, it does not improve the performance of existing algorithms on datasets where the data is already roughly aligned and scaled. Our method has the potential to be extended to more complex geometric relationships among local features, as we illustrate in the experiments.

1. Introduction

Automatic visual classification and object recognition promise to make computer vision a key component in applications such as surveillance, computer interaction, data mining, assistance for the visually impaired. The problem is difficult because the same object or category can manifest itself in a variety of ways due to "nuisances" such as changes of viewpoint, visibility, illumination and clutter, in addition to inter-class variability.

The effect of such nuisances can be mitigated by designing image statistics (a.k.a. "features") that are insensitive to them, later to be fed to a classifier, or could be "learned away" by a super-classifier given sufficient training data. For instance, while illumination variability cannot be eliminated [5], it can be mitigated by using gradient orientation [2] rather than image intensity statistics, as now customary in popular features such as SIFT [21]. Geometric nuisances include viewpoint variations and visibility effects, such as occlusions and cast shadows. The former can be approximated locally by affine image domain deformations, while the latter cannot be eliminated by design, which has prompted many to restrict the domain of features to small regions (a.k.a. "patches" or "local features"), while deferring to the classifier the choice of which features belong to the "object" and which to the "background."

Indeed, image deformations due to viewpoint variations could be eliminated altogether, not just for locally planar scenes, but this comes at the cost of discarding all geometric information [38]. This result gives theoretical grounding to so-called "bags-of-features" approaches to visual recognition [39], where local features are compared regardless of their position, and partially explains their striking success that took many in the community by surprise.

However, the results in [38] assume that objects of any shape are equally likely, and therefore a true invariant feature has to "normalize" all possible shapes. This is not the case in reality, where natural scenes exhibit significant regularity in their geometry. This suggests that geometric information may be important and should be exploited in recognition systems, and idea that has recently been put to fruition, as we discuss in the next subsection.[1] Because modeling geometric variability explicitly is rather complex, existing approaches either restrict the attention to simple models that are too stiff, such as affine procrustes ([10] and references therein), or try to impose some topological or lose geometric constraints to otherwise geometry-free rep-

*The work was conducted when the first author was with University of California Los Angeles.

[1] We believe that the importance of geometry has been underestimated, in part because popular datasets (e.g. Caltech 101) exhibit artificially limited variability with objects roughly centered and scaled, as illustrated by [12].

resentations such as bags of features *a-posteriori*.

We propose to use the *proximity distribution of vector-quantized local image statistics as a global image descriptor* that captures both photometric and geometric information. We then propose a kernel, called *Proximity Distribution Kernel (PDK)*, that can be readily used to design a classifier. Features and classifiers are two aspects of the same problem, and we believe that they should be designed jointly. We do so, and demonstrate our approach on every challenging experiment that we have tried, including standard datasets such as Graz-01 and Graz-02 [28] and Pascal [6]. Our approach does not improve the state of the art on the Caltech 101 dataset [7], for obvious reasons[1], although it is within 10% of the current best algorithm.

1.1. Related work

Bags-of-features or -words have shown remarkable performance in recognition of objects and categories [39, 41]. In light of [38] one could guess that the best performers would not be the ones using local features with the highest possible level of invariance, a fact shown eloquently in [41], which confirmed previous observations [19]. Nevertheless, recent attempts to enforce loose spatial information have shown great promise, including [11] where features are augmented with their spatial coordinates in the pyramid matching kernel (PMK). A different method to exploit geometric constraint was proposed by [16], who used "semi-local parts" that combine neighboring features by validating their affine relations, and later [17] framed them into a maximum-entropy approach that performed well on textures as well as objects. Even more recently, and more directly related to our manuscript, [19] proposed using spatial pyramid matching for recognizing natural scene categories, using dense SIFT-like descriptors at regular grid points. In this case, performance on Caltech 101 registered at 64.6%, but did not perform as well on the Graz dataset, purportedly because of the large geometric variability.

One of the first attempts to impose topological constraints by joining features into pairs is [34], whereas [40] used proximity, measured by the Euclidean distance between feature coordinates, to improve their bag-of-words algorithm for testing on the Caltech101. This exploits the nature of the dataset, where foreground images are roughly aligned. In [26] proximity is used to learn compositional categorization models. Features were organized into triplets by [36], who used the order type index histogram as a qualitative image descriptor. Performance on Caltech 6 improves [10] on two of the four categories in the data set. In [33] the joint statistics of local neighborhood operators are used for object recognition and localization. Mercer kernels were applied to semi-local groups of features by [22], where features are grouped by nearest neighbors, whereas [24] developed a spatial weighting technique that assigns low weight to background features. The proposed method has been shown to improve the traditional bag-of-words on the PASCAL challenge [6] that is a four-category classification task. Local geometric information is also represented in a template-based approach by [3], who introduced the notion of geometric blur. Second-order geometric relationships is further used [4] to solve the correspondence between geometric blur features, and then applied it to category recognition problems. Small collections of points were also used by [8] using triangulated polygons compared via their log-anisotropy, whereas pair-wise constraints were exploited in [29] for human detection. In [31] multiple segmentations encode lose spatial structure, whereas [23] use Delaunay triangulation to determine the neighborhood structure that is important to fix the scale of textons. Context modeling has recently potential to improve object detection [14]. Hierarchical structure can also helps to improve local features; examples can be found in [1] and [25]. Most recently, Leordeanu et al. [20] also proposed using pairwise configurations between edge fragments for category recognition. Recognition task is formulated as a quadratic assignment problem and model parameters are learned sequentially. Compared to [20], our approach is much simpler in that no object model is required.

The most related work is the *correlogram* first proposed in [15] and later extended by Savarese et al. [32]. In [32], a correlagram is used to measure the distribution of distances between all pairs of visual labels and then applied to category classification tasks (with combination of label distribution). In comparison, PDK captures rank information between visual words, which is more reliable and sparse. In addition, our method is simpler since it works without combining other representations. It is interesting to compare and/or combine these two approaches in the future.

Our manuscript introduces a novel representation that simultaneously encodes local photometric information (from vector-quantized local feature descriptors) and local geometric information (from proximity distributions), and proposes a matched classifier based on a kernel defined on such distributions. We introduce our approach next.

2. Sorting Out Bags of Features with Spatial Information

Let $I : D \subset \mathbb{R}^2 \to \mathbb{R}^3$ be an image and $\{\phi_k\}_{k=1,\dots,M}$ a collection of M local features extracted from I: $\phi : \{I(x),\ x \in \Omega \subset D\} \to \mathbb{R}^K$, for instance [21, 13]. On the set of all available features, we perform vector-quantization, to arrive at V codebook elements $\mathcal{V} = \{v_1, \dots, v_V\}$, each in $\mathbb{R}^K$, together with the position of the centroid of each feature $x_k = \int_{\Omega_k} x dx / \int_{\Omega_k} dx$ where Ω_k is the support region of the k-th feature. So, in this first coding stage, the image I is represented by the *local feature* $\{(x_k, \alpha_k)\}_{k=1,\dots,M}$ with

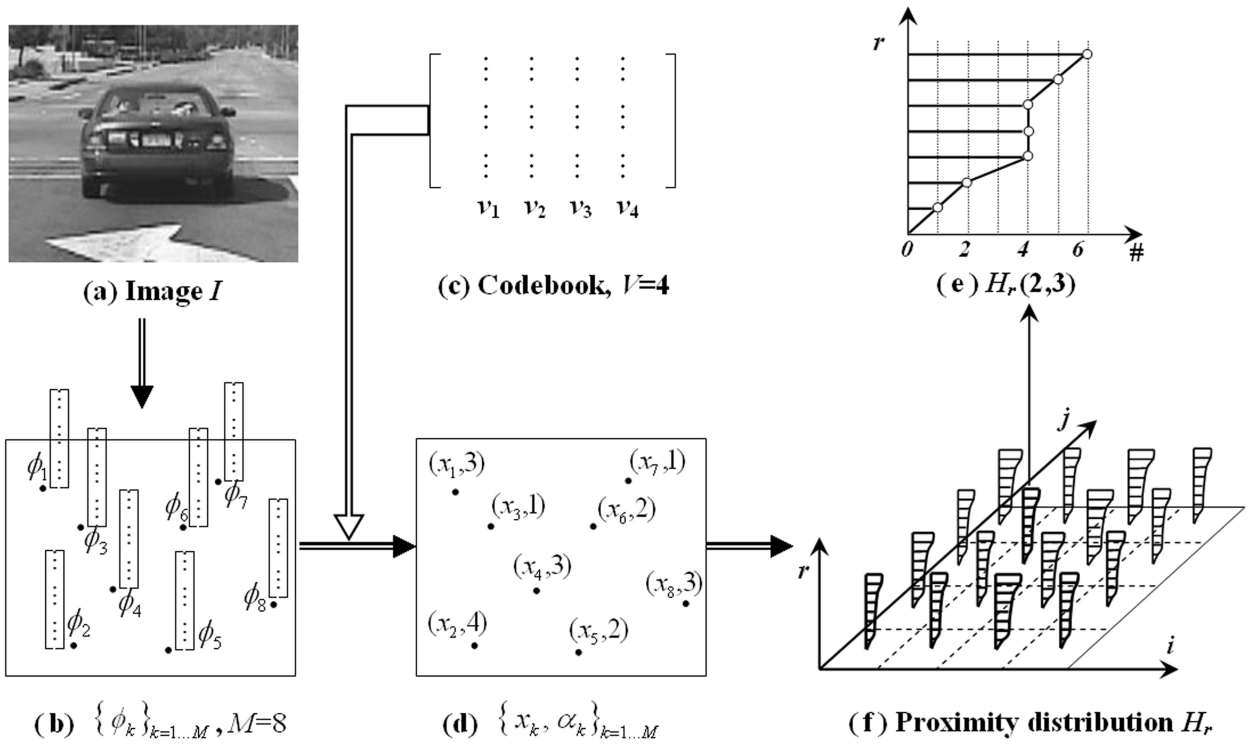

Figure 1. Demonstration of building proximity distribution of local features. (a) An input image I. (b) Local features. (c) Codebook with size $V = 4$. (d) Local words and their positions. (e) An example cumulative proximity distribution $H_r(2, 3)$ for word pair v_2, v_3. (f) The array of proximity distributions $H_r(i, j)$.

each α_k identified with the integers $k = 1, \dots, V$. An illustration of this representations is shown in Fig. 1 (a-d). In the next subsection we introduce the proximity distribution of these local features.

2.1. Proximity Distributions of Local Features

Given an image I, and its associated coding $\{(x_k, \alpha_k)\}_{k=1,\dots,M}$, we now construct a two-dimensional array of one-dimensional proximity distributions by considering, for each codeword pair v_i, v_j, the number of features $H_r(i, j)$ of type j that are within r-nearest neighbors of a feature of type i.

$$H_r(i,j) = \#\{(\alpha_l, \alpha_m) : \alpha_l = i, \alpha_m = j, \\ d_{NN}(x_l, x_m) \le r\}, \quad r = 1, \dots, n_r \tag{1}$$

where $d_{NN}(x_l, x_m) \le r$ indicates that x_m is within the r-th nearest neighbors of x_l (strictly speaking, this depends on the set $\{x_k\}_{k=1,\dots,M}$). n_r is the size of the neighborhood of interest. Note that this defines a one-dimensional array (indexed by r) for fixed $i, j = 1, \dots, V$, corresponding to an un-normalized cumulative probability distribution. This includes local photometric information, encoded in the codewords returned by vector-quantization, as well as local geometric information, encoded by the proximity cumulogram. Examples of such distributions are shown in Fig. 1 (e-f).

Another choice is to use the histogram instead of the cumulative distribution, i.e., $d_{NN} = r$ instead of $d_{NN} \le r$ in (1). This may sound more natural; however, in practice the histogram is often less stable. On the other hand, the cumulative distribution is better tailored for our purpose: For example, the $\mathbb{L}^1$ norm between two cumulative (one-dimensional) histograms is equivalent to the Earth Mover distance [30] (or Wasserstein, Ornstein & Mallows distance) between the them. Other advantages of this representation will become apparent shortly.

2.2. Proximity Distribution Kernel

Now that we have a representation of the images, we need to introduce a method to compare them. As mentioned above, we can use the $\mathbb{L}^1$ distance to build the extended Gaussian kernel. However, because of visibility artifacts, as we discuss further, it is better to use a "minimum" kernel for this purpose. Given two images I^1 and I^2, represented by their distributions H_r^1, H_r^2, we define the *Proximity Distribution Kernel* (PDK) as

$$K(I^1, I^2) \doteq \sum_{i,j=1}^{V} \sum_{r=1}^{k} \min(H_r^1(i,j), H_r^2(i,j)). \tag{2}$$

This distance measures the similarity of proximity distribution, or co-occurrence of codewords in close spatial proximity. The use of the minimum, already introduced by [11], affords some resistance to clutter without the need for more rigid models that account for occlusions explicitly in a hypothesis testing scenario [10]. Naturally, local spatial deformations are lost in the distribution: Although the nearest-neighbor relationships are discontinuous with respect to domain deformations, the use of the cumulative distribution smooths out the effects of such discontinuities (a similar outcome could have been obtained by taking the spatial density and smoothing it).

One very important property of PDK is that it satisfies the Mercer's condition, i.e., it is a positive semi-definite kernel. This is clear from the fact that $\min(.,.)$ is a Mercer kernel and that the set of Mercer kernels is closed under summation. This property guarantees consistency in a reproducing-kernel Hilbert space scenario, and is best suited for use with a support vector machine (SVM) [37] for classification, as we articulate in our experimental section.

2.3. Extensions beyond second-order statistics

The construction above could be easily extended to proximity relationships between more than two features, for instance to triplets arranged into cubic arrays, and to higher-order as well. However, we have found empirically that even simple pairs of features represent a good tradeoff between capturing some spatial relationships without overly constraining the representation. For the purpose of illustration, we outline the construction for triplets, leaving the extension to higher-order statistics to the reader.

The geometry of a triplet of features is characterized by the triangle formed from their coordinates. To be invariant to similarity transformation, we represent a triangle by the so called two-dimensional Bookstein coordinates, as used in [9]. Given an image I and its coded local feature sets $\{(x_k, \alpha_k)\}_{k=1,\ldots,M}$, the *triplet proximity distribution* is captured as a three-dimensional array of two-dimensional histograms, $T_{t_1,t_2}(i,j,k)$, where t_1 and t_2 are discrete Bookstein coordinates. In other words, a triplet proximity distribution T is a five-dimensional histogram that measures the joint distribution of word triplets and triangular shapes. Similar to PDK, the similarity between two distributions (T^1, T^2) can be computed by the following *triplet-PDK*

$$K_T(I^1, I^2) \doteq \sum_{i,j,k=1}^{V} \sum_{t_1,t_2} \min(T^1_{t_1,t_2}(i,j,k), T^2_{t_1,t_2}(i,j,k)). \tag{3}$$

The triplet-PDK is essentially a histogram intersection [35] for five-dimensional histograms. It is obvious that the Triplet-PDK also satisfies Mercer's condition and therefore also suitable to the kernel-based approaches.

2.4. Implementation Issues

To capture the co-occurrence distribution of k features requires a computational complexity of $O(M^k)$. Specifically, the complexity to build $H_r(i,j)$ is $O(M^2)$ and to build $T_{t_1,t_2}(i,j,k)$ is $O(M^3)$. In practice, however, due to the sparseness of local features and therefore the co-occurrence joint histograms, it is often more efficient to compute PDK (for pairs or triplets) directly without explicitly computation of H or T. For example, in formulae (2) and (3), the summation over all (i,j) or (i,j,k) are often not necessary. This is particularly useful when one is interested in capturing higher-order of co-occurrences, for instance between quadruplets of features.

One parameter for PDK is the size of neighborhood n_r. Theoretically, large n_r will capture more information until saturation occurs (e.g., for $n_r > M-1$). This is consistent with our observation in our preliminary experiments. However, the computational complexity and memory requirements are linearly increased with n_r. In practice the performance is not very sensitive for $n_r \geq 64$ for images with several hundreds of features. We discuss the actual choice of numerical parameters in the next section.

3. Experiments

We test the proposed approach on several public image data sets, including the Graz-01, Graz-01, and the PASCAL challenge. Only grayscale images are used in all the experiments, although there may be rooms for further improvement by including color information.

Our experiments mainly focus on the pairwise proximity case, i.e. PDK. The triplet case or the triplet-PDK is tested in one experiment for the sake of comparison with PDK.

3.1. Graz-01

The Graz-01 data set [27] contains two object classes containing 373 bike images and 460 person images. In addition, it has a background class with 270 images. The data set is challenging because of the large variability in object scale, pose and illumination. Some example images are shown in Fig. 2.

Two experiments, called "test I" and "test II," have been performed on this data set. In both tests, we use $n_r = 200$ for PDK.

In test I [27, 41], specific training and testing sets are used as done in [28], which contains 200 training images and 100 testing images for each category. Table 1 shows our experimental result for this test along with previously reported scores. From the table it is clear that PDK outperforms existing methods. It is worth noting that a kernel-

Figure 2. Example images from Graz-01. Top row: bike images. Middle row: person images. Bottom row: background images.

based approach (EMD kernel with SVM) is also used by [41], without taking into account any spatial relations. In addition, both SIFT and spin image [18] are used in [41], while only SIFT are used in our experiment.

In test II [19], for each non-background class, 100 positive and 100 negative images (50 from the background class and 50 from the other class) are used for training. Testing is on a similarly distributed set but the number of images are reduced by half. The result is an averaged equal-error rate over ten runs. Table 2 shows our experimental results for test II. In addition to PDK, we also tested the triplet-PDK with only 300 randomly sampled features for each image for computational efficiency.

Two important lessons were learned in test II. First, it shows that our method is more robust in dealing with large geometric deformations when compared to the pyramid matching technique (PMK) used in [19]. It is worth noting that PMK has demonstrated excellent performance when images are roughly aligned, such as those in Caltech 101. In other words, the comparison with PMK verifies both the importance of including spatial information and the importance of been robust for large image deformations - both are taken care of by our approaches. Second, although complex geometric relations captured by triplet-PDK may improve the performance, the simple pairwise relation captured by standard PDK represents a good tradeoff between insensitivity to geometric deformation and preservation of discriminative power of the representation.

Figure 3. Incorrectly classified images in test I. Top left: a bike image classified as no bike. Top right: a background image classified as a bike image. Bottom left: a person image classified as no person. Bottom right: a background image classified as a person image.

Figure 4. Incorrectly classified images in test II. Top left: a bike image classified as no bike. Top right: a person image classified as no person. Middle left: a person image classified as a bike image. Middle right: a bike image classified as a person image. Bottom left: a background image classified as a bike image. Bottom right: a background image classified as a person image.

3.2. Graz-02

The Graz-02 data set [28] is an improved version of the Graz-01 data set and designed to be even more challenging.

Table 1. Results (EER %) of test I on the Graz-01 data set [27].

Class	Boosting [27] +SIFT	SVM [41]+ (SIFT+Spin)	Pair. Inter.[20]	PDK+ SIFT
Bikes	86.5	92.0	84.0	**95.0**
Person	80.8	**88.0**	82.0	**88.0**
Ave.	83.7	90.0	83.0	**91.5**

Table 2. Results (EER) of test II on the Graz-01 data set [27].

Class	PMK [19]	T-PDK (300 samp)	PDK
Bikes	86.3±2.5	88.9±1.8	**90.2±2.6**
Person	82.3±3.1	85.1±3.1	**87.2±3.8**

Table 3. Recognition results (rate at EER %) on Graz-02 data set [28]. In the fifth column, only 600 randomly sampled features per image are used for computational efficiency.

	Boost.+ SIFT[28]	Boost.+ Comb.[28]	Pair. Inter.[20]	PDK+ SIFT	PDK+ hybrid
Bike	76.0	77.8	**92.0**	86.7	86.0
Person	70.0	81.2	86.0	86.7	**87.3**
Car	68.9	70.5	n/a	**74.7**	**74.7**
Ave.	71.6	76.5	n/a	**82.7**	**82.7**

It contains three object classes including 365 bike images, 311 person images, 420 car images, and 380 counter-class images. In addition, the Graz-02 has been carefully balanced with respect to backgrounds for all categories. We follow the experimental setup of [28]. More specifically, for each non-background class, 150 positive and 150 negative (counter-class) images are used for training. Similarly, 75 positive and 75 negative images are used for testing. Some example images are shown in Fig. 5.

Opelt et al. [28] presented a framework using boosting to learn a subset of feature vectors and combine them into final hypothesis for category classification. They test their method with several different types of feature descriptors on the Graz-02 data set. The best recognition rate of their method was achieved by using a combination of different features. In our experiment, similar to [28], two versions of feature descriptors are used. The first one uses only standard SIFT features (i.e. blobs only), which corresponding to the fourth column in Table 3. The second one uses both corners and blob regions same as in our experiment for Graz-01, which corresponding to the fifth column in Table 3. The experimental results are listed in Table 3. From the table it is clear that our method outperforms previously proposed approaches for both standard SIFT and hybrid features.

3.3. PASCAL 2005

We further tested our approaches on the PASCAL VOC Challenge [6]. This challenge contains both classification and detection tasks. In this paper we focus on the former. The classification task consists of binary classification on four categories: motorbikes, bikes, person, and cars. Some

Figure 5. Example images from Graz-02. Each row contains two images from one category, from top to bottom: bikes, person, cars, background.

example images are shown in Fig. 6. The classification contains an easy test (test 1) and a difficult test (test 2). We applied our approach to the difficult one because there is not much to be improved for the easy one. Eleven teams attended the challenge for test 2, the best score is achieved by Zhang and Schmid using bag-of-features (SIFT) with an extended Gaussian kernel.

Similar to Zhang and Schmid, we use SIFT descriptors on scale invariant corners and blobs. We use a vocabulary set with 200 words and set $n_r = 64$ for our proposed kernel. The experiments is listed in Table 4 together with the best score achieved in the PASCAL challenge[2]. From the table we see that our approach outperforms the PASCAL winner in half of the four categories and in the average performance. While examining the dataset more closely, we observed that images from the categories "motorbike" and "bike" have more extreme clutter than those from the cat-

[2]Higher scores (except for cars) were achieved in [24] using prior (manual) segmentation during training process.

Figure 6. Example images from PASCAL 2005. Each row contains two images from one category, from top to bottom: motorbikes, bikes, person, cars. Note that some images contains multiple categories.

Table 4. Testing on PASCAL challenging [6] (test 2).

	Motor	Bike	Person	Car	Average
Winner [6]	**79.8**	**72.8**	71.9	72.0	74.1
PDK	76.9	70.1	**72.5**	**78.4**	**74.5**

egories "person" and "car." This is one reason why our method works so much better on "car" images than on "motorbikes."

4. Discussion

Our goal in this work is to bring geometric context information back into category recognition, after the surprising success of bag-of-words approaches. We have shown that the proximity distribution of vector-quantized features represents an effective compromise between achieving insensitivity to geometric deformations, for instance due to viewpoint variations in addition to inter-class shape variability, and maintaining discriminative power, so that geometric structure, albeit in "weak form," is taken into account.

The proximity distribution kernel (PDK) has been designed with this goal in mind, and we have shown that it can be easily computed for pairwise relationships, and can be easily extended to capturing higher-order co-occurrence statistics, hence capturing the local geometric context of an image. Insensitivity to occlusions is achieved through the use of a "min" functional in the definition of the kernel, and insensitivity to the discontinuities in the proximity relations is achieved through the use of the probability distribution in place of a (density) histogram.

We have demonstrated empirically the potential of our approach, by showing that it outperforms all existing ones in scenarios where there is significant geometric variability, such as in the Graz datasets, and in the PASCAL challenge. Other datasets where geometric variability is limited, e.g. Caltech 101, do not exalt our approach, since the data is provided in (roughly) aligned and scaled form to begin with.

Our approach opens the door to additional ways to include spatial structure into image representations for recognition, including higher-order statistics (involving more than two features at a time), as we have illustrated. Other uses of our representation, for instance for object detection, are certainly possible and will be the subject of future work.

Acknowledgement

We thank Andrea Vedaldi for helpful discussion. This work was supported by ONR N00014-03-1-0850/67F-1080868 and AFOSR FA9550-06-1-0138.

References

[1] A. Agarwal and B. Triggs. "Hyperfeatures: Multilevel local coding for visual recognition". *ECCV*, I:30-43, 2006. 2

[2] L. Alvarez, F. Guichard, P. L. Lions, and J. M. Morel. "Axioms and fundamental equations of image processing", Springer Berlin / Heidelberg, 1993. 1

[3] A. C. Berg and J. Malik, "Geometric Blur for Template Matching", *CVPR*, I:607-614, 2001. 2

[4] A. C. Berg, T. L. Berg, and J. Malik. "Shape Matching and Object Recognition Using Low Distortion Correspondences". *CVPR*, I:26-33, 2005. 2

[5] H. Chen, P. Belhumeur and D. W. Jacobs. "In search of Illumination Invariants", *CVPR*, 254-261, 2000. 1

[6] M. Everingham, A. Zisserman, C. K. I. Williams, L. Van Gool, et al. "The 2005 PASCAL Visual Object Classes Challenge". In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognis-*

ing Textual Entailment, eds. J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d'Alche-Buc, LNAI 3944, 117-176, Springer-Verlag, 2006. 2, 6, 7

[7] L. Fei-Fei, R. Fergus, and P. Perona. "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories".*CVPR Workshop on Generative-Model Based Vision*, 2004. 2

[8] P. F. Felzenszwalb. "Representation and Detection of Deformable Shapes", *PAMI*, 27(2):208-220, 2005. 2

[9] P. Felzenszwalb and J. Schwartz. "Hierarchical Matching of Deformable Shapes", *CVPR*, 2007. 4

[10] R. Fergus, P. Perona and A. Zisserman. "Object Class Recognition by Unsupervised Scale-Invariant Learning", *CVPR*, II:264-271, 2003. 1, 2, 4

[11] K. Grauman and T. Darrell. "The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features". *ICCV*, II:1458-1465, 2005. 2, 4

[12] G. Griffin, A. D. Holub, and P. Perona. "The Caltech-256", Caltech Technical Report. 1

[13] C. Harris and M. Stephens, "A combined corner and edge detector", *Alvey Vision Conference*, 147-151, 1988. 2

[14] D. Hoiem, A. A. Efros, and M. Hebert. "Putting Objects in Perspective ". *CVPR*, II:2137-2144, 2006. 2

[15] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih. "Color-spatial Indexing and Applications", *IJCV*, 35(3):245-268, 1999. 2

[16] S. Lazebnik, C. Schmid, and J. Ponce. "Semi-Local Affine Parts for Object Recognition". *BMVC*, 2:959-968, 2004. 2

[17] S. Lazebnik, C. Schmid, and J. Ponce. "A Maximum Entropy Framework for Part-Based Texture and Object Recognition". *ICCV*, 1:832-838, 2005. 2

[18] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using affine-invariant regions," *PAMI*, 27(8):1265-1278, 2005. 5

[19] S. Lazebnik, C. Schmid, and J. Ponce. "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, II:2169-2178, 2006. 2, 5, 6

[20] M. Leordeanu, M. Hebert, and R. Sukthankar. "Beyond Local Appearance: Category Recognition from Pairwise Interactions of Simple Features", *CVPR*, 2007. 2, 6

[21] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *IJCV*, 60(2):91-110, 2004. 1, 2

[22] S. Lyu, "Mercer Kernels for Object Recognition with Local Features", *CVPR*, 2005. 2

[23] J. Malik, S. Belongie, T. Leung, and J. Shi. "Contour and texture analysis for image segmentation". *IJCV*, 43(1):7-27, 2001. 2

[24] M. Marszalek and C. Schmid. "Spatial Weighting for Bag-of-Features", *CVPR*, II:2118- 2125, 2006. 2, 6

[25] J. Mutch and D. G. Lowe. "Multiclass Object Recognition with Sparse, Localized Features". *CVPR*, I:11-18, 2006. 2

[26] B. Ommer and J. M. Buhmann. "Learning Compositional Categorization Models", *ECCV*, III:316-329, 2006. 2

[27] A. Opelt, M. Fussenegger, A. Pinz and P. Auer. "Weak Hypotheses and Boosting for Generic Object Detection and Recognition", *ECCV*, 2004. 4, 6

[28] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. "Generic Object Recognition with Boosting", *PAMI*, 28(3), 2006. 2, 4, 5, 6

[29] X. Ren, A. C. Berg, and J. Malik. "Recovering Human Body Configurations Using Pairwise Constraints between Parts". *ICCV*, 824-831, 2005. 2

[30] Y. Rubner, C. Tomasi, and L. J. Guibas. "The Earth Mover's Distance as a Metric for Image Retrieval", *IJCV*, 40(2):99-121, 2000. 3

[31] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. "Using Multiple Segmentations to Discover Objects and their Extent in Image Collections", *CVPR*, II:1605-1614, 2006. 2

[32] S. Savarese, J. M. Winn, and A. Criminisi. "Discriminative Object Class Models of Appearance and Shape by Correlatons". *CVPR*, II:2033-2040, 2006. 2

[33] B. Schiele and A. Pentland. "Probabilistic Object Recognition and Localization. In *ICCV*, 177-182, 1999. 2

[34] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. "Discovering Objects and their location in images", *ICCV*, 2005. 2

[35] M. J. Swain and D. H. Ballard. "Color Indexing", *IJCV*, 7(1):11-32, 1991. 4

[36] J. Thureson and S. Carlsson."Appearance Based Qualitative Image Description for Object Class Recognition", *ECCV*, 518-529, 2004. 2

[37] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995. 4

[38] A. Vedaldi and S. Soatto, "Features for Recognition: Viewpoint Invariance for Non-Planar Scenes". *ICCV*, II:1474-1481, 2005. 1, 2

[39] J. Willamowski, D. Arregui, G. Csurka, C. Dance, and L. Fan. "Categorizing Nine Visual Classes using Local Appearance Descriptors", *ICPR Workshop Learning for Adaptable Visual Systems*, 2004. 1, 2

[40] H. Zhang, A. Berg, M. Maire, and J. Malik. "SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition". *CVPR*, 2006. 2

[41] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. "Local Features and Kernels for Classifcation of Texture and Object Categories: A Comprehensive Study." *IJCV*, 73(2):213-238, 2007. 2, 4, 5, 6