Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study

Jianguo Zhang, Marcin Marszalek, Svetlana Lazebnik, Cordelia Schmid

Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study

Jianguo Zhang[†] Marcin Marszałek[†] Svetlana Lazebnik[‡] Cordelia Schmid[†]

[†]*INRIA Rhône-Alpes, LEAR - GRAVIR*
655 av. de l'Europe, 38330 Montbonnot, France
zhang@inrialpes.fr marszale@inrialpes.fr

[‡]*Beckman Institute, University of Illinois*
405 N. Mathews Ave., Urbana, IL 61801, USA
slazebni@uiuc.edu schmid@inrialpes.fr

Abstract

Recently, methods based on local image features have shown promise for texture and object recognition tasks. This paper presents a large-scale evaluation of an approach that represents images as distributions (signatures or histograms) of features extracted from a sparse set of keypoint locations and learns a Support Vector Machine classifier with kernels based on two effective measures for comparing distributions, the Earth Mover's Distance and the χ^2 distance. We first evaluate the performance of our approach with different keypoint detectors and descriptors, as well as different kernels and classifiers. We then conduct a comparative evaluation with several state-of-the-art recognition methods on 4 texture and 5 object databases. On most of these databases, our implementation exceeds the best reported results and achieves comparable performance on the rest. Finally, we investigate the influence of background correlations on recognition performance.

1. Introduction

The recognition of texture and object categories is one of the most challenging problems in computer vision, especially in the presence of intra-class variation, clutter, occlusion, and pose changes. Recent achievements in both texture and object recognition have demonstrated that using local features, or descriptors computed at a sparse set of scale- or affine-invariant keypoints, tends to be an effective approach [7, 11, 20]. At the same time, Support Vector Machine (SVM) classifiers [18] have shown their promise for visual classification tasks, and the development of specialized kernels suitable for use with local features has emerged as a fruitful line of research [8]. To date, most evaluations of methods combining kernels and local features have been small-scale and limited to one or two datasets. This motivates us to build an effective image classification approach combining a bag-of-keypoints representation with a kernel-based learning method and to test the limits of its performance on the most challenging databases available today.

Our study consists of three components:

Evaluation of implementation choices. We assess many alternative implementation choices, including keypoint detector type, level of geometric invariance, feature descriptor, and classifier kernel. This evaluation yields several insights of practical importance. For example, a combination of multiple detectors and descriptors usually achieves better results than even the most discriminative individual detector/descriptor channel. Also, for most datasets in our evaluation, local features with the highest possible level of invariance do not yield the best performance.

Comparison with existing methods. We conduct a comparative evaluation with several state-of-the-art methods for texture and object classification on 4 texture and 5 object databases. For texture classification, our approach outperforms existing methods on Brodatz [2], KTH-TIPS [9] and UIUCTex [11] datasets and obtains comparable results on the CUReT dataset [3]. For object category classification, our approach outperforms existing methods on Xerox7 [20], Graz [16], CalTech6 [7], CalTech101 [6] and the more difficult test set of the Pascal challenge [5]. It obtains comparable results on the easier Pascal test set. The power of orderless bag-of-keypoints representations may be not surprising in the case of texture images, which lack clutter and have uniform statistical properties. However, it is not a priori obvious that such representations are sufficient for object category classification, since they ignore spatial relations and do not separate foreground from background features.

Influence of background features. In many existing datasets, background features tend to be correlated with the foreground (e.g., cars are often pictured on a road, while faces appear in office environments). Since our bag-of-keypoints method uses both foreground and background features to classify the image, it is important to investigate whether background features provide any "hints" for recognition. Using a novel methodology, we study the influence of background features on the diverse and challenging Pascal benchmark. Our experiments reveal that while background does contain some discriminative information for the foreground category, using foreground and back-

ground features together *does not* improve the performance of our method. Thus, even in the presence of background correlations, the features on the objects themselves are the key to recognition. Moreover, our experiments also show the danger of using monotonous or highly correlated backgrounds for training, since this leads to poor recognition performance on test sets with more complex backgrounds.

We have deliberately limited our evaluations to the image-level *classification* task, i.e., classifying an entire test image as containing an instance of one of a fixed number of given object classes. This task must be clearly distinguished from *localization*, i.e., reporting a location hypothesis for an object. We demonstrate that, given the right implementation choices, simple orderless image representations can be surprisingly effective on a wide variety of imagery. Thus, they can serve as good baselines for measuring the difficulty of newly acquired datasets and for evaluating more sophisticated recognition approaches that incorporate structural information about the object.

The rest of this paper is organized as follows. Relevant previous work on texture and object recognition is discussed in Section 1.1. The components of our approach (keypoint detectors, descriptors, and classifier kernels) are described in Section 2. Section 3.1 describes our experimental setup and the nine databases included in the current study. Section 3.2 evaluates the implementation choices relevant to our approach. Sections 3.3 and 3.4 present comparisons with existing texture and object category classification methods, and Section 4 evaluates the effect of background correlations on recognition performance. Finally, Section 5 concludes the paper with a summary of our findings and future extensions.

1.1. Related Work

Texture recognition. Recently, there has been a great deal of interest in recognizing images of textured surfaces subjected to lighting and viewpoint changes [3, 11, 19]. The basic idea of these methods is to represent texture images as distributions or *texton histograms* over a universal *texton dictionary*. Several texton-based representations using features derived from filter bank outputs and raw pixel values were developed [19], and further improvements were achieved by using SVM classifier with a kernel based on χ^2 histogram distance [9]. However, a major shortcoming of these methods is that the underlying representation is not geometrically invariant. No adaptation is performed to compensate for changes in scale or surface orientation with respect to the camera. By contrast, Lazebnik et al. [11] have proposed an intrinsically invariant representation based on distributions of appearance descriptors computed at a *sparse* set of affine-invariant keypoints (as opposed to earlier *dense* approaches that compute descriptors at every pixel). We take this approach as a starting point and further improve its discriminative power with the help of a kernel-based learning method.

Object recognition. The earliest work on appearance-based object recognition has utilized *global* descriptions such as color or texture histograms. The main drawback of such methods is their sensitivity to clutter and occlusions. For this reason, global methods were gradually supplanted by *part-based methods*, e.g. [7], that combine appearance descriptors of local features with a representation of their spatial relations. While part-based models offer an intellectually satisfying way of representing objects, learning and inference problems for spatial relations are complex and computationally intensive, especially in a *weakly supervised* setting where the location of the object in a training image has not been marked. On the other hand, orderless bag-of-keypoints methods [20] have the advantage of simplicity and computational efficiency, though they fail to represent the geometric structure of the object or to distinguish between foreground and background features. For these reasons, bag-of-keypoints methods can be adversely affected by clutter, just as earlier global methods. To overcome this problem, novel SVM kernels that can yield high discriminative power despite noise and clutter have been proposed recently [8]. While these methods have obtained promising results, they have not been extensively tested on databases with heavily cluttered, uncorrelated backgrounds, so the true extent of their robustness has not been conclusively determined. Our own approach is related to that of Grauman and Darrell [8], who have developed a kernel that approximates the optimal partial matching between two feature sets. Specifically, we use a kernel based on the *Earth Mover's Distance* [17], i.e. the exact partial matching cost.

2. Components of the representation

2.1. Image representation

We use two complementary local region detector types to extract salient image structures: The *Harris-Laplace* detector [15] responds to corner-like regions, while the *Laplacian* detector [12] extracts blob-like regions. These two detectors are invariant to scale transformations only – they output circular regions at a certain characteristic scale. We obtain rotation invariance by rotating the regions in the direction of the dominant gradient orientation [15] and affine invariance through the use of an *affine adaptation* procedure [15]. Affinely adapted detectors output ellipse-shaped regions which are then *normalized*, i.e., transformed into circles. In summary, our detectors offer different levels of invariance: scale invariance only (S), scale with rotation invariance (SR), and affine invariance (A). We denote the Harris detector with different levels of invariance as HS, HSR and HA and the Laplacian detector as LS, LSR and LA.

The normalized circular patches obtained by the detec-

tors serve as domains of support for computing appearance-based descriptors. We use the SIFT descriptor [13] which computes a gradient orientation histogram within the support region and the SPIN descriptor [11] which is a rotation-invariant two-dimensional histogram of intensities within an image region.

We consider each detector/descriptor pair as a separate "channel." The combination of multiple detector/descriptor channels is denoted by (detector + detector) (descriptor + descriptor), e.g., (HS+LS)(SIFT+SPIN) means the combination of HS and LS detectors each described with SIFT and SPIN descriptors.

2.2. Comparing distributions of local features

To compare sets of local features, we represent their distributions in the training and test images. One method for doing this is to cluster the set of descriptors found in each image to form its *signature* $\{(p_1, u_1), \ldots, (p_m, u_m)\}$, where m is the number of clusters, p_i is the center of the ith cluster, and u_i is the proportional size of the cluster. We extract 40 clusters with k-means for each image. *Earth Mover's Distance* (EMD) [17] has shown to be very suitable for measuring the similarity between image signatures. The EMD between two signatures $S_1 = \{(p_1, u_1), \ldots, (p_m, u_m)\}$ and $S_2 = \{(q_1, w_1), \ldots, (q_n, w_n)\}$ is defined as $D(S_1, S_2) = \left[\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\, d(p_i, q_j)\right] / \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}$, where f_{ij} is a flow value that can be determined by solving a linear programming problem, and $d(p_i, q_j)$ is the *ground distance* between cluster centers p_i and q_j. We use Euclidean distance as the ground distance.

An alternative to image signatures is to obtain a global *texton vocabulary* (or *visual vocabulary*) by clustering descriptors from a training set, and then to represent each image as a histogram of texton labels [19, 20]. Given a global texton vocabulary of size m, the ith entry of a histogram is the proportion of all descriptors in the image having label i. To compare two histograms $S_1 = (u_1, \ldots, u_m)$ and $S_2 = (w_1, \ldots, w_m)$, we use the χ^2 distance defined as $D(S_1, S_2) = \frac{1}{2}\sum_{i=1}^{m}\left[(u_i - w_i)^2/(u_i + w_i)\right]$.

2.3. Kernel-based classification

For classification, we use *Support Vector Machines* (SVM) [18]. For a two-class problem the decision function has the form $g(x) = \sum_i \alpha_i y_i K(x_i, x) - b$, where $K(x_i, x)$ is the value of a *kernel function* for the training sample x_i and the test sample x. The $y_i \in \{-1, +1\}$ and α_i are the class label and the learned weight of the training sample x_i. b is a learned threshold parameter. The training samples with $\alpha_i > 0$ are usually called *support vectors*.

We use the two-class setting for binary detection, i.e., classifying images as containing or not a given object class. For multi-class classification, we use the one-against-one technique, which trains a classifier for each possible pair of classes. When classifying an image, we evaluate all binary classifiers, and perform voting [18].

To incorporate EMD or χ^2 distance into the SVM framework, we use generalized Gaussian kernels $K(S_i, S_j) = \exp\left(-D(S_i, S_j)/A\right)$ where $D(S_i, S_j)$ is EMD (resp. χ^2 distance) if S_i and S_j are image signatures (resp. vocabulary histograms). The resulting kernel is the *EMD kernel* (or χ^2 *kernel*). The parameter A of the EMD (resp. χ^2) kernel is the mean value of the EMD (resp. χ^2) distances between all training images. To combine channels, we apply the generalized Gaussian kernel to the summed distance $D = \sum_i^n D_i$, where D_i is the distance for channel i.

3. Empirical Evaluation

3.1. Experimental setup and datasets

For our experimental evaluation, we use 4 texture and 5 object category datasets, described below. For texture classification, we randomly select 100 different training/test splits and report the average classification accuracy, together with the standard deviation, over the 100 runs. For object classification, we use the same training and test sets as the publications with which we are comparing.

The UIUCTex dataset [11] (Fig. 1) contains 25 texture

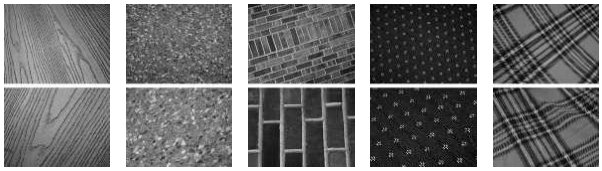

Figure 1. Image examples of the UIUCTex dataset.

classes with 40 images per class. Textures are viewed under significant scale and viewpoint changes. Non-rigid deformations, illumination changes and viewpoint-dependent appearance variations are also present. The KTH-TIPS dataset [9] (Fig. 4) contains 10 texture classes with 81 im-

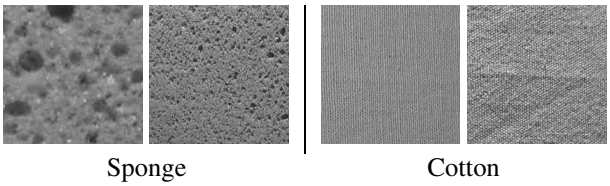

Figure 4. Image examples of the KTH-TIPS database.

ages per class. Images are captured at nine scales spanning two octaves, viewed under different illumination directions and different poses. The Brodatz texture album [2] contains 112 different texture classes where each class is represented by one image divided into nine sub-images. Note that this dataset is somewhat limited, as it does not model viewpoint, scale, or illumination changes. For the CUReT tex-

Figure 2. Categories of CalTech101 with the best classification rates (left) and with the lowest rates (right).

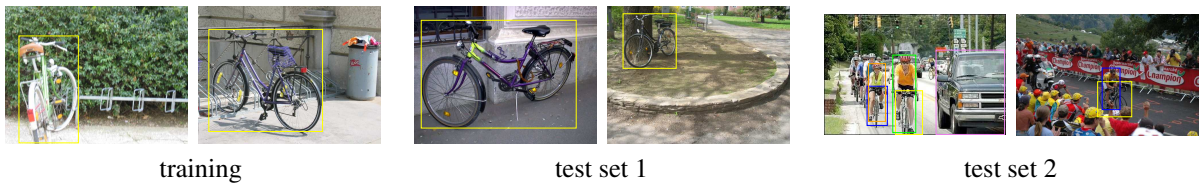

Figure 3. Image examples with ground truth object annotation of the bike categories of the Pascal challenge.

ture database [3] we use the same subset of images as [19]. This subset contains 61 texture classes with 92 images for each class. These images are captured under different illuminations with seven different viewing directions. The changes of viewpoint, and, to a greater extent, of the illumination direction, significantly affect the texture appearance.

The Xerox7 dataset [20] (Fig. 5) consists of of seven

Figure 5. Correctly recognized (first row) and misclassified (second row) Xerox7 images.

classes. It includes images with highly variable pose and background clutter, and the intra-class variability is large. We perform multi-class classification with the same setup as in [20]. The CalTech6 database [7] contains six classes and a background set. We use the same training and test set for two-class classification (object vs. background) as [7]. The Graz dataset [16] (Fig. 6) contains persons, bikes and a

Figure 6. Recognized and missed images of the Graz dataset.

background class. We use the same training and test set for two-class classification as [16]. The CalTech101 dataset [6] (Fig. 2) contains 101 object categories. Most of the images in the database contain little or no clutter. Furthermore, the objects tend to lie in the center of the image and to be present in similar poses. Some images have a partially black background due to artificial image rotations. We follow the experimental setup of Grauman et al. [8], i.e., we randomly select 30 training images per class and test on the remaining images, reporting the average accuracy for all the classes. The Pascal dataset [5] (Fig. 3) includes bicycles, cars, motorbikes and people. It has one training dataset and two test sets. In the "easier" test set 1, images are taken from the same distribution as the training images. In the "harder" test set 2, images are collected by Google search. An additional complication is that many images in test set 2 contain instances of several classes.

3.2. Evaluation of parameters

Evaluation of different levels of invariance. First, we show the results of evaluating different levels of invariance (S, SR, A) of our two keypoint detectors on several datasets. The number of training images per class are 20 for UIUC-Tex, 3 for Brodatz, 100 for Graz bikes. For Xerox7, we use tenfold cross-validation. These settings are kept for all experiments reported in this section. In this test, all regions are described with the SIFT descriptor and the EMD kernel is used for classification. Table 1 shows that pure scale invariance (S) performs best for the Brodatz, Graz bikes and

Databases	Scale Inv.	Scale and Rotation	Affine Inv.
	HS+LS	HSR+LSR	HA+LA
UIUCTex	92.2 $\pm$1.4	98.0 $\pm$0.5	98.0 $\pm$0.6
Brodatz	94.4 $\pm$0.7	94.0 $\pm$0.9	91.3 $\pm$1.1
Graz bikes	91.9 $\pm$2.6	91.3 $\pm$2.6	90.5 $\pm$3.0
Xerox7	94.7 $\pm$1.2	92.2 $\pm$2.3	91.4 $\pm$1.8

Table 1. Evaluation of different levels of invariance.

Xerox7 datasets, while for UIUCTex, rotation invariance (SR) is important. The reason is that Brodatz, Graz and Xerox7 have no rotation or affine changes, while UIUCTex has significant viewpoint changes and arbitrary rotations. Even in this case, affine-invariant features fail to outperform the scale- and rotation-invariant ones. The apparent superiority of scale-invariant detectors for recognition could be due to their greater robustness, as the affine adaptation process can often be unstable in the presence of large affine or perspective distortions.

Evaluation of different channels. Next, we compare the performance of different detector/descriptor channels and their combinations. We use the EMD kernel for classification. Table 2 shows results obtained for the UIUCTex dataset. We can see that the Laplacian detector tends to perform better than the Harris detector. The most likely reason for this difference is that the Laplacian detector usually extracts four to five times more regions per image than Harris-Laplace, thus producing a richer representation. Using the two detectors together tends to further raise performance. SIFT performs slightly better than SPIN. Combining SIFT with SPIN boosts the overall performance because the two descriptors capture different kinds of information (gradients vs. intensity values). These observations are confirmed by results on other datasets [21] (omitted here for lack of space). Overall, the combination of Harris-Laplace and Laplacian detectors with SIFT and SPIN is the preferable choice in terms of classification accuracy, and this is the setup used in Sections 3.3 and 3.4.

Channels	SIFT	SPIN	SIFT+SPIN
HSR	97.1 $\pm$ 0.6	93.9 $\pm$ 1.1	97.4 $\pm$ 0.6
LSR	97.7 $\pm$ 0.6	93.9 $\pm$ 1.0	98.2 $\pm$ 0.6
HSR+LSR	98.0 $\pm$ 0.5	96.2 $\pm$ 0.8	98.3 $\pm$ 0.5

Table 2. Detector and descriptor evaluation on UIUCTex.

Evaluation of different kernels. The learning ability of a kernel classifier depends on the type of kernel used. Here we compare SVM with three different kernels: linear, χ^2, and EMD. As a baseline, we also evaluate EMD with nearest neighbor (NN) classification – the same setup as in Lazebnik et al. [11]. For the signature-based classifiers (EMD-NN and EMD kernel), we use 40 clusters per image as before. For the other SVM kernels, which work on histogram representations, we create a global vocabulary by concatenating 10 clusters per class. For UIUCTex, Brodatz, Graz bikes and Xerox7, the vocabulary sizes are 250, 1120, 20 and 70, respectively. Table 3 shows classification results for the LSR+SIFT channel, which are representative of all other channels. We can see that EMD-NN always performs worse than the EMD kernel, i.e., that a discriminative approach gives a significant improvement. The difference is particularly large for the Xerox7 database, which has wide intra-class variability. Among the vocabulary/histogram representations, the χ^2 kernel performs better than linear. The performance levels of EMD kernel and the χ^2 kernel are comparable and either of them is a good choice in our framework provided that a suitable vocabulary can be built efficiently. To avoid the computational expense of building global vocabularies for each dataset, we use the EMD kernel in the following experiments.

Databases	Vocabulary-Histogram		Signature	
	Linear	χ^2	EMD-NN	EMD
UIUCTex	97.0 $\pm$ 0.6	98.1 $\pm$ 0.6	95.0 $\pm$ 0.8	97.7 $\pm$ 0.6
Brodatz	96.1 $\pm$ 0.8	96.0 $\pm$ 0.7	86.5 $\pm$ 1.2	94.1 $\pm$ 0.8
Graz bikes	83.9 $\pm$ 3.6	83.8 $\pm$ 2.0	84.6 $\pm$ 3.4	89.8 $\pm$ 2.6
Xerox7	79.8 $\pm$ 3.0	89.2 $\pm$ 2.1	59.4 $\pm$ 4.1	92.4 $\pm$ 1.7

Table 3. Classification accuracy of different kernels.

3.3. Texture classification

In this section, we present a comparative evaluation of our approach with four state-of-the-art texture classification methods: Lazebnik [11], VZ-joint [19], Hayman [9] and global Gabor as in [14]. Table 5 shows the classification accuracy of the five different methods for 4 texture databases. We use (HS+LS)(SIFT+SPIN) as image description for all databases except for UIUC, for which we use the rotation invariant version. For the **UIUCTex** database we can observe that both our method and Lazebnik's method work much better than Hayman's method and VZ-joint, while Hayman's method works better than VZ-joint. Overall, the improved performance of our method over Lazebnik's and of Hayman over VZ-joint shows that discriminative learning helps to achieve robustness to intra-class variability. On this dataset, global Gabor features perform the worst, since they are not invariant and averaging the features over all pixels loses discriminative information. Overall, the three non-invariant dense methods in our evaluation have relatively weak performance on this database. For the **KTH-TIPS** database we can observe that our method works best, Hayman's comes second, and VZ-joint and Lazebnik's method are below them. Global Gabor filters come last, though they still give good results and their performance is significantly higher for this database than for UIUCTex. This may be due to the relative homogeneity of the KTH-TIPS texture classes. For the **Brodatz** our method performs best, closely followed by Hayman's method. We can see that Hayman's method performs better than VZ-joint, and our method better than Lazebnik's method, i.e. that kernel-based learning

Training set size	UIUCTex 20	KTH-TIPS 40	Brodatz 3	CUReT 43
ours	98.3 $\pm$ 0.5	95.5 $\pm$ 1.3	95.4 $\pm$ 0.3	95.3 $\pm$ 0.4
Hayman	92.0 $\pm$ 1.3	94.8 $\pm$ 1.2	95.0 $\pm$ 0.8	98.6 $\pm$ 0.2
Lazebnik	96.4 $\pm$ 0.9	91.3 $\pm$ 1.4	89.8 $\pm$ 1.0	72.5 $\pm$ 0.7
VZ-joint	78.4 $\pm$ 2.0	92.4 $\pm$ 2.1	92.9 $\pm$ 0.8	96.0 $\pm$ 0.4
G. Gabor	65.2 $\pm$ 2.0	90.0 $\pm$ 2.0	87.9 $\pm$ 1.0	92.4 $\pm$ 0.5

Table 5. Comparison of different methods for texture datasets.

Methods	Xerox7	CalTech6	Graz	Pascal test set 1	Pascal test set 2	CalTech101
(HS+LS)(SIFT+SPIN)	94.3	97.9	90.0	92.8	74.3	53.9
other	82.0 [20]	96.6 [20]	83.7 [16]	94.6 [10]	70.5 [4]	43 [8]

Table 4. Comparison with the best reported results on several object datasets.

improves the performance over 1-NN classification. For the **CUReT** database Hayman's method obtains the best results, followed by VZ-joint, our method, global Gabor filters, and Lazebnik's method. On this dataset, local feature methods are at a disadvantage. Since most of the CUReT textures are very homogeneous and high-frequency, lacking salient structures such as blobs and corners, keypoint extraction does not produce very good image representations. A simple patch descriptor seems to be more appropriate.

In conclusion, our method achieves the highest accuracy on three texture databases and comparable results on the CUReT dataset. Its robustness to viewpoint and scale changes has been clearly demonstrated on the UIUCTex and the KTH-TIPS datasets. Our results show that for most datasets, combining geometric invariance at the representation level with a discriminative classifier at the learning level results in a very effective system. Note that even though impressive results are obtained using VZ-joint (patch descriptors) on the CUReT and Brodatz datasets, this method does not perform as well on the other datasets, thus showing its limited applicability. An important factor affecting the performance of local feature methods is image resolution, since keypoint extraction tends to not work well on low-resolution images. This concurs with the previous results showing the advantage of the denser Laplacian detector over the sparser Harris-Laplace.

3.4. Object category classification

In this section we evaluate our approach for object category classification and compare it to the best results reported in the literature. Table 4 shows the results for different datasets. The EMD kernel and SVM are used. For details on the experimental setup see section 3.1.

Results for multi-class classification on **Xerox7** show that our method outperforms the Xerox bag-of-keypoints method [20] in the same experimental setting. This is due to the fact that we use a combination of detectors and descriptors, a more robust kernel (see table 3) and scale invariance as opposed to affine invariance (see table 1). Fig. 5 shows a few results on Xerox7. Two-class classification (object vs. background) accuracy on the **CalTech6** and **Graz** databases is reported with the ROC equal error rate. Our approach outperforms existing methods on both datasets. Note that results for CalTech6 are high, indicating the relatively low level of difficulty of this dataset. Fig. 6 shows some results for the Graz datasets. Misclassified bikes are either observed from the front, very small, or only partially visible. Misclassified people are either observed from the back, occluded, or very small. We also evaluate our approach for the object category classification task of the **Pascal** challenge [5]. Table 4 shows ROC equal error rates of our method for detecting each class vs. the other best method reported in the Pascal challenge. For test set 1 the best results, slightly better than ours, were obtained by Jurie and Triggs [10]. This approach uses a dense set of multi-scale patches instead of a sparse set of descriptors computed at interest points. For test set 2 best results, below ours, were obtained by Deselaers et al. [4]. They use a combination of patches around interest points and patches on a fixed grid. Results for multi-class classification on **Caltech101** show that our approach outperforms Grauman et al. [8] for the same setup. The best results on this dataset (48%) are currently reported by Berg et al. [1]. However, these results are not comparable to ours, since they were obtained in a supervised setting with manually segmented training images. Fig. 2 presents the categories with the best and worst classification rates. We can observe that some of the lowest rates are obtained for categories that are characterized by their shape as opposed to texture, such as anchors.

To conclude, our method achieves the best results on Xerox7, CalTech6, Graz, Pascal test set 2 and CalTech101 and is comparable for Pascal test set 1.

4. Object categories – influence of background

Our method recognizes object categories taking both foreground and background features as input. In most databases included in our evaluation, object categories have fairly characteristic backgrounds, e.g., most car images contain a street or a parking lot. In this section, our goal is to determine whether background correlations provide our method with additional cues for classification. In the following experiments, we use the (HS+LS)(SIFT) channels – SPIN is dropped for computational efficiency – and the EMD kernel. The signature size is set to 40 per image. Our test bed is the Pascal database, which comes with complete ground truth annotations (see Fig. 3). Using the annotations, we extract two sets of features from each image: foreground features (FF) that are within the ground truth object region, and background features (BF) that are outside. Throughout our experiments, we systematically replace the original background features from an image by two specially constructed alternative sets: *random* and *constant natural scene* backgrounds (referred to as BF-RAND and BF-CONST, respectively). BF-RAND are obtained by randomly shuffling background features among all of the images in the dataset. For example, the background of a face image may be re-

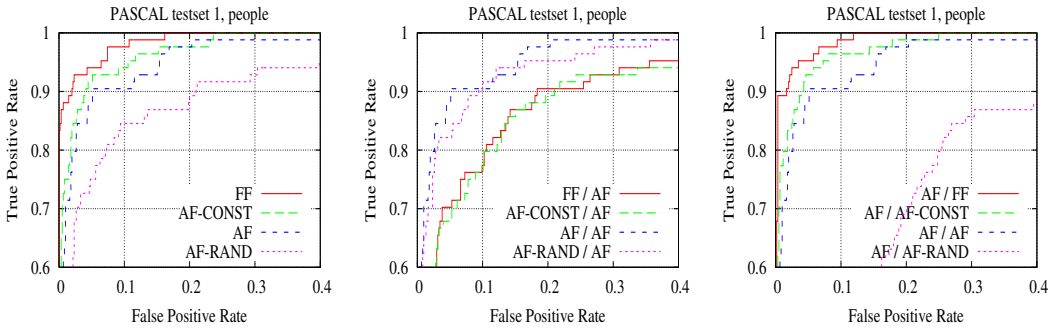

Figure 7. ROC curves for training and testing with four combinations of the foreground features with different types of background.

placed by the background of a car image. BF-CONST are background features extracted from images captured by a fixed camera observing a natural scene over an extended period of time, so they include continuous lighting changes and the movement of trees and clouds.

Fig. 8 shows ROC curves obtained by training and testing on only the background features (BF). Judging by the shape of the ROC curves, the background features contain a lot of discriminative information for test set 1, and significantly less for test set 2 (e.g., the performance of background features for test set 2 bicycles is close to chance level). The performance curves of the BF-RAND and BF-CONST feature sets (not shown in the figure) are at chance level as one would expect, since BF-RAND and BF-CONST do not contain any information about the foreground object by construction.

Fig. 7 evaluates combinations of foreground features with different types of background features. Due to space limitations only results for the people test set 1 are presented. Results for the other test sets are similar [21]. AF denotes all the features extracted from the original image, AF-RAND denotes the combination of FF and BF-RAND and AF-CONST denotes the combination of FF and BF-CONST. Fig. 7 (left) shows ROC curves for a situation where training and testing are performed on the same feature combination. FF always gives the highest results, indicating that object features play the key role for recognition, and recognition with segmented images achieves better performance than without segmentation. Mixing background features with foreground features *does not* give higher recognition rates than FF alone. For images with roughly constant backgrounds (AF-CONST), the performance is almost the same as for images with foreground features only. It is intuitively obvious that classifying images with fixed backgrounds is as easy as classifying images with no background clutter at all. Finally, the ROC curves for AF-RAND are the lowest, which shows that objects with uncorrelated backgrounds are harder to recognize. Fig. 7 (middle) shows ROC curves for a setup where the training set has different types of backgrounds and the test set has its original background. We can observe that training on AF or AF-RAND while testing on AF gives the highest results. Thus, even under randomly changed training backgrounds, the SVM can find decision boundaries that generalize well to the original training set. Training on FF or AF-CONST and testing on AF gives lower results, most likely because the lack of clutter in FF set and the monotonous backgrounds in AF-CONST cause the SVM to overfit the training set. By contrast, varying the object background during training, even by random shuffling, seems to prevent this. Finally, Fig. 7 (right) shows ROC curves for a situation where the training set has the original backgrounds and the test set has different types of backgrounds. When the test set is "easier" than the training one, performance improves, and when it is "harder," the performance drastically drops. This is consistent with the results of Fig. 7 (middle), where training on the "harder" sets AF or AF-RAND gave much better results than training on the "easier" sets FF and AF-CONST.

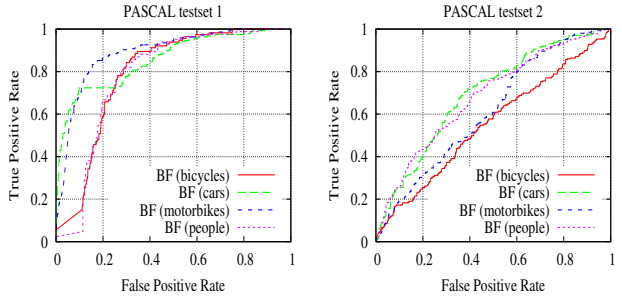

Figure 8. ROC curves of object classification obtained by training and testing on background features only.

Based on our evaluation of the role of background features in bag-of-keypoints classification, we can venture two general observations. First, while the backgrounds in most available datasets have non-negligible correlations with the foreground objects, using both foreground and background features for learning and recognition does not result in better performance for our method. In our experimental setting, the recognition problem is easier in the absence of clutter. Second, when the statistics of the test set are unknown at training time, it is usually beneficial to pick the most difficult training set available.

5. Discussion

In this paper, we have investigated the performance of an image classification approach combining a bag-of-keypoints representation with a kernel-based learning method. Results on challenging datasets have shown that surprisingly high levels of performance can be achieved not only on texture images, which are clutter-free and relatively statistically homogeneous, but also on object images containing substantial clutter and intra-class variation.

One of the contributions of our paper is a comprehensive evaluation of multiple keypoint detector types, levels of geometric invariance, feature descriptors and classifier kernels. This evaluation has revealed several general trends. We show that to achieve the best possible performance, it is necessary to use a combination of several detectors and descriptors together with a classifier that can make effective use of the complementary types of information contained in them. Also, we show that using local features with the highest possible level of invariance usually does not yield the best performance. Thus, a practical recognition system should seek to incorporate multiple types of complementary features, as long as their local invariance properties do not exceed the level absolutely required for a given application.

In testing our method on 4 texture and 5 object databases, we have followed an evaluation regime far more rigorous than that of most other comparable works. In fact, our evaluation of multiple texture recognition methods highlights the danger of the currently widespread practice of developing and testing a recognition method with only one or two databases in mind. For example, methods tuned to achieve high performance on the CUReT database (e.g., the VZ method) have weaker performance on other texture databases, such as UIUCTex, and vice versa, methods tuned to UIUCTex and Brodatz (e.g., the Lazebnik method) perform poorly on CUReT.

Another contribution of our paper is our evaluation of the influence of background features. It shows the pitfalls of training on datasets with uncluttered or highly correlated backgrounds, since this yields disappointing results on test sets with more complex backgrounds.

Future research should focus on designing more effective feature detectors and descriptors, for example for representing shape, as well as designing kernels that incorporate geometrical relations between local features. In the longer term, successful category-level object recognition and localization is likely to require more sophisticated models that capture the 3D shape of real-world object categories as well as their appearance. In the development of such models and in the collection of new datasets, simpler bag-of-keypoints methods can serve as effective baselines and calibration tools.

Acknowledgments

This research was supported by the French ACI project MoViStaR, the UIUC-CNRS-INRIA collaboration agreement and the European projects LAVA and PASCAL. M. Marszałek was supported by the INRIA student exchange program and a grant from the European Community under the Marie-Curie project VISITOR. Svetlana Lazebnik was supported by the National Science Foundation under grants IIS-0308087 and IIS-0535152.

References

[1] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *CVPR*, 2005.

[2] P. Brodatz. *Textures: A Photographic Album for Artists and Designers*. Dover, New York, 1966.

[3] K. Dana, B. van Ginneken, S. Nayar, and J. Koenderink. Reflectance and texture of real world surfaces. *ACM Transactions on Graphics*, 18(1), 1999.

[4] T. Deselaers, D. Keysers, and H. Ney. Discriminative training for object recognition using image patches. In *CVPR*, 2005.

[5] M. Everingham, A. Zisserman, C. Williams, L. V. Gool, et al. The 2005 PASCAL visual object classes challenge. In *First PASCAL Challenge Workshop*.

[6] 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. In *CVPR Workshop on Generative-Model Based Vision*, 2004.

[7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.

[8] K. Grauman and T. Darrell. Pyramid match kernels: Discriminative classification with sets of image features. In *ICCV*, 2005.

[9] E. Hayman, B. Caputo, M. Fritz, and J.-O. Eklundh. On the significance of real-world conditions for material classification. In *ECCV*, 2004.

[10] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, 2005.

[11] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *PAMI*, 27(8), 2005.

[12] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2), 1998.

[13] D. Lowe. Distinctive image features form scale-invariant keypoints. *IJCV*, 60(2), 2004.

[14] B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *PAMI*, 18(5), 1996.

[15] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1), 2004.

[16] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *ECCV*, 2004.

[17] Y. Rubner, C. Tomasi, and L. Guibas. The Earth Mover's distance as a metric for image retrieval. *IJCV*, 40(2), 2000.

[18] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Cambridge, MA, 2002.

[19] M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *CVPR*, 2003.

[20] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *IWLAVS*, 2004.

[21] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classifcation of texture and object categories: An in-depth study. Technical Report RR-5737, INRIA Rhône-Alpes, 2005.