Improving a Discriminative Approach to Object Recognition using Image Patches

Thomas Deselaers, Daniel Keysers, and Hermann Ney

Lehrstuhl für Informatik VI – Computer Science Department,
RWTH Aachen University – D-52056 Aachen, Germany
{deselaers, keysers, ney}@informatik.rwth-aachen.de

Abstract. In this paper we extend a method that uses image patch histograms and discriminative training to recognize objects in cluttered scenes. The method generalizes and performs well for different tasks, e.g. for radiograph recognition and recognition of objects in cluttered scenes. Here, we further investigate this approach and propose several extensions. Most importantly, the method is substantially improved by adding multi-scale features so that it better accounts for objects of different sizes. Other extensions tested include the use of Sobel features, the generalization of histograms, a method to account for varying image brightness in the PCA domain, and SVMs for classification. The results are improved significantly, i.e. on average we have a 59% relative reduction of the error rate and we are able to obtain a new best error rate of 1.1% on the Caltech motorbikes task.

1 Introduction

The learning of representations that allow for recognition and classification of objects in cluttered scenes is a significant open problem in computer vision. A very promising approach to this problem assumes that the objects to be learned and recognized consist of a collection of parts, and that different objects can share some of the parts. Additionally, changes in the geometrical relation between image parts can be modeled to be flexible to tolerate some deformations. This approach can handle occlusions well, because if some parts of an object are occluded, the other parts can still be detected and recognized perfectly.

Related work includes Mohan and colleagues [13] who use predetermined parts of human bodies to detect humans in cluttered scenes. Dorko and Schmid [4] use image patches to classify cars, but the extracted patches from the training set are labelled by whether they are part of a car or not. Leibe and Schiele [10] use scale-invariant interest points and manually segmented training data for classification. In contrast to these approaches, we need only weak supervision in training, i.e. only information about the presence of an object in the image. Fergus and colleagues [5] and Weber and colleagues [14] statistically model position, occurrence, and appearance of object parts.

In [3] we present and compare several approaches to use histograms of image patches for the recognition of objects in cluttered scenes. It was shown that an

approach using histograms of vector quantized image patches and discriminative training performed best among the tested approaches (global patch search & direct voting, nearest neighbor, naive Bayes, generative Gaussians, and discriminative training). In this work, we further investigate this approach and propose several extensions: 1. As proposed in [9], we use image patches in various scales, enabling us to account for objects at different scales. 2. We use Sobel filtered images in addition to the gray values to account for edge structures in the images. 3. As the histograms created are very sparse (e.g. there are approx. 1000 data points in a 4096 bin histogram), we generalize the histograms to use non-binary bin assignments. 4. To account for different lighting conditions, we incorporate a method for brightness normalization.

2 Baseline Approach

The method for discriminative training of image patch histograms, which has been proposed in [3], consists of two steps: 1. feature extraction and 2. training and classification. These steps are briefly summarized in the following sections.

2.1 Feature Extraction

Given an image, we use up to 500 square image patches as features. These patches are extracted around interest points obtained using the method proposed by Loupias and colleagues [11]. Additionally, we use 300 patches from a uniform grid of 15×20 cells that is projected into the image. In contrast to the interest points from the detector, these points can also fall onto very homogeneous areas of the image. This property is important for capturing homogeneity in objects in addition to points that are detected by interest point detectors, which are usually of high variance. Figure 1 shows the points of interest detected in a typical image. The patches are allowed to extend beyond the image border, in which case the part of the patch falling outside the image is padded with zeroes. After the patches are extracted, a PCA dimensionality reduction is applied to reduce the large dimensionality of the data, keeping 40 coefficients. These data are then clustered with a Linde-Buzo-Gray algorithm using the Euclidean distance. Then we discard all information for each patch except its closest corresponding cluster center identifier. For the test data, this identifier is determined by evaluating the Euclidean distance to all cluster centers for each patch. Thus, the clustering assigns a cluster $c(x) \in \{1, \dots C\}$ to each image patch x and allows us to create histograms of cluster frequencies by counting how many of the extracted patches belong to each of the clusters. The histogram representation $h(X)$ with C bins is then determined by counting and normalization such that $h_c(X) = \frac{1}{L_X} \sum_{l=1}^{L_X} \delta(c, c(x_l))$, where δ denotes the Kronecker delta function, $c(x_l)$ is the closest cluster center to x_l, and x_l is the l-th image patch extracted from image X.

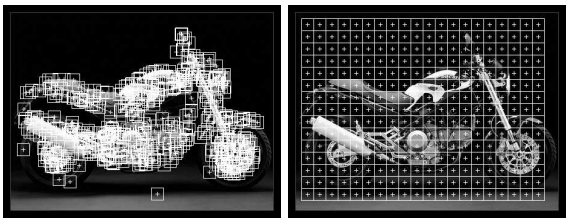

Fig. 1. Patch extraction: salient points and uniform grid.

2.2 Decision Rule

Having obtained this representation by histograms of image patches, we define a decision rule for the classification of images. In [3] we observed that the method using discriminative training of log-linear models outperforms other methods.

The approach based on maximum likelihood of the class-conditional distributions does not take into account the information of competing classes during training. We can use this information by maximizing the class posterior probability $\prod_{k=1}^{K}\prod_{n=1}^{N_k} p(k|X_{kn})$ instead. Assuming a Gaussian density with pooled covariances for the class-conditional distribution, this maximization is equivalent to maximizing the parameters of a log-linear or maximum entropy model

$$p(k|h) = \frac{1}{Z(h)} \exp\left(\alpha_k + \sum_{c=1}^{C} \lambda_{kc} h_c\right),$$

where $Z(h) = \sum_{k=1}^{K} \exp\left(\alpha_k + \sum_{c=1}^{C} \lambda_{kc} h_c\right)$ is the renormalization factor. (Note that also the generative Gaussian model can be rewritten in this form. Furthermore, we can always find a generative model that results in the same posterior distribution [7].) The maximizing distribution is unique and the resulting model is also the model of highest entropy with fixed marginal distributions of the features [7]. Efficient algorithms to determine the parameters $\{\alpha_k, \lambda_{kc}\}$ exist. We use a modified version of generalized iterative scaling [1]. Bayes' decision rule is used for classification.

3 Extensions to the method

In [3], we thoroughly investigated several decision rules and classification methods. In this work, we further investigate the feature extraction method and test several extensions.

Multi-scale features. In the original approach, all patches extracted were of the same size and we have experimentally evaluated which image patch size performed best on the given tasks. This can lead to problems if the objects to be recognized are of different scales. Here we propose to extract patches of different sizes. That is, at each feature extraction point, we extract square patches of 7, 11, 21, and 31 pixels width. To be able to use these patches in the proposed training

and classification framework, all extracted patches are scaled to a common size of 15×15 pixels using the Bresenham scaling algorithm.

Derivatives. In many applications of pattern recognition, derivatives can improve classification performance significantly, e.g. in automatic speech recognition, derivatives are normally used. Also in the recognition of handwritten characters, derivatives can strongly improve the results [8], as the local derivatives allow mapping of edges to edges. To take advantage of these effects, we enrich the patches by their horizontally and vertically Sobel filtered versions. That is, the data is tripled by adding horizontally and vertically Sobel-filtered patches. Then, the PCA transformation is applied to all three versions (gray values, horizontal Sobel, vertical Sobel) at once and the dimensionality is reduced from $3 \cdot 15^2 = 675$ to 40 in total to allow for efficient processing in the remaining steps (clustering and histogramization).

Histogram smoothing. A weakness of the original approach might be that the histograms are high dimensional and very sparse, e.g. the histograms have 4096 bins but only 800 patches (2400 for multi-scale features) are extracted per image. Thus, most of the bins are empty and cannot contribute to the result. To have smoother histograms, we generalize the histograms to use non-binary bin assignments, i.e. patches do not only contribute to their closest cluster center but to all cluster centers that are sufficiently similar. That is, given an image patch and the Euclidean distance $d_i := d(x, c_i)$ to cluster center c_i, the corresponding histogram count h_i is updated as

$$h_i \leftarrow h_i + \frac{\exp(-\frac{d_i}{\alpha})}{\sum_{i'} \exp(-\frac{d_{i'}}{\alpha})}.$$

By changing α, the strength of smoothing can be changed.

Brightness normalization. Another issue which is a well-known problem in computer vision is that different images are often taken under different lighting conditions, and thus the brightness of otherwise very similar images can vary significantly. Evidently, the brightness of an image should usually not change class membership, but e.g. the Euclidean distance between two images that are identical except for their brightness can be very high. A practical approach for brightness normalization in this context is given in the PCA transformation: The first PCA vector for a collection of image patches usually captures the change in brightness and thus contributes most to the overall brightness of the image patches. Thus we propose to discard the first component of the PCA transformed vectors in order to discard information about global brightness of image patches [12]. Figure 2 illustrates this effect: For each of the three tasks (airplanes, faces, motorbikes), it shows the first component of the PCA matrix (clearly capturing global patch brightness) and an example of a bright and a dark patch reconstructed from the PCA transformed and dimensionality reduced representations, one with and one without the first PCA component. It can be observed that the differences in brightness are reduced for the patches which have been reconstructed without the first PCA component.

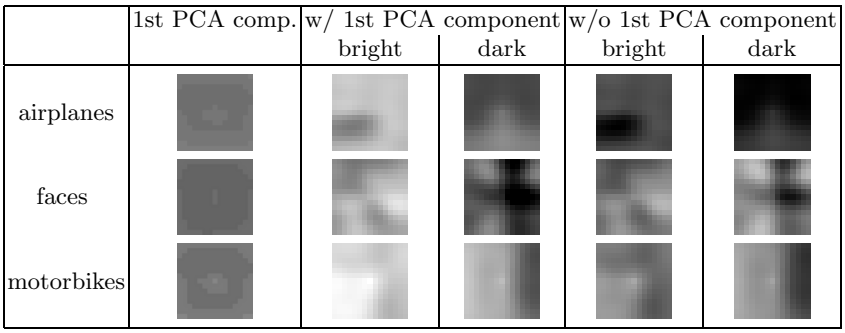

Fig. 2. For each of the three tasks: First PCA component, a bright image patch and a dark image patch reconstructed from the PCA vectors using all 40 PCA components, and the same patches reconstructed from the PCA vectors discarding the first PCA component.

Support-Vector-Classifier. As support-vector-machines (SVM) are a known to be a good classification method, the maximum likelihood approach was exchanged in favor of a support-vector classified from libsvm[1]. We tried radial basis functions and linear functions as kernels and optimized the parameters using the training data.

4 Databases

Fergus and colleagues [5] use different datasets for unsupervised object training and recognition of objects. The task is to determine whether an object is present in an image or not. For this purpose, several sets of images containing certain objects (airplanes, faces, and motorbikes) and a set of background images not containing any of these objects are available[2], which we use in the experiments. The images are of various sizes, and for the experiments they were converted to gray images. The airplanes and the motorbikes task consists of 800 training and 800 test images each. The faces task consists of 436 training and 434 test images. For each of these tasks, half of the images contain the object of interest and the other half does not. An example image of each set is shown in Figure 3. For our experiments we scaled all images to a common height of 225 pixels as our approach implicitly learns the importance of the image size for classification otherwise [3].

5 Experimental Results

In Table 1 we give the results for the baseline method from [3] and the results obtained with the proposed extensions in comparison to results from the literature. All experiments were carried out using 4096 dimensional histograms.

[1] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[2] `http://www.robots.ox.ac.uk/~vgg/data`

Fig. 3. Examples from the Caltech database (airplanes, faces, motorbikes, background).

Table 1. Summary of results and comparison to results from other publications.

Method		airplanes	faces	motorbikes
Discriminative Model	[3]	2.6	5.8	1.5
+ multi-scale features		**1.1**	5.0	1.9
+ multi-scale & Sobel features		4.5	13.6	2.6
+ multi-scale feat. & fuzzy hist.		2.6	8.1	1.4
+ multi-scale & brightness norm.		1.4	**3.7**	**1.1**
lin. SVM + multi-scale & brightness norm.		2.4	7.8	2.1
rbf. SVM + multi-scale & brightness norm.		2.1	9.4	2.1
Statistical Model	[5]	9.8	3.6	7.5
Texture features	[2]	**0.8**	1.6	7.4
Segmentation	[6]	2.2	**0.1**	10.4

Comparing the results using multi-scale features to the results from the baseline method where only patches of one size were extracted, a clear improvement can be seen in two of the three tasks. The result for the motorbikes task was not improved. These results can be explained by the fact that the scale of the motorbikes is very homogeneous and thus, multi-scale features cannot improve the results. Due to the positive results, all experiments in the following were performed using multi scale features.

The results where Sobel features were used are worse than those from the baseline method. This unexpected result may be due to the combined PCA transformation of brightness and contrast information. We will further investigate the reasons for these effects.

In a next step, we evaluated the possible advantages of fuzzy histograms. Figure 4 shows the effect of choosing different parameters α to smooth the image patch histograms. In these experiments we used 4096 clusters and multi scale features. The figures show that the fuzzy histograms do not improve the results in this setting. In Figure 5 we compare fuzzy histograms with discrete histograms using different numbers of histogram bins. It can be seen that fuzzy histograms outperform discrete histograms in the case of only few clusters. As the clustering process is computationally very expensive, but the creation of fuzzy histograms is not more expensive than the creation of discrete histograms given a cluster model, fuzzy histograms can be used to obtain reasonable results when computing power for the training is limited. It can also be clearly seen that the number of clusters has less impact on the classification performance when

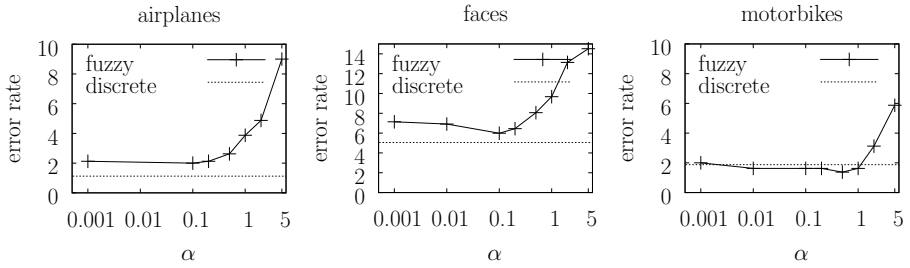

Fig. 4. Error rates for the Caltech tasks depending on the smoothing factor α in fuzzy histograms.

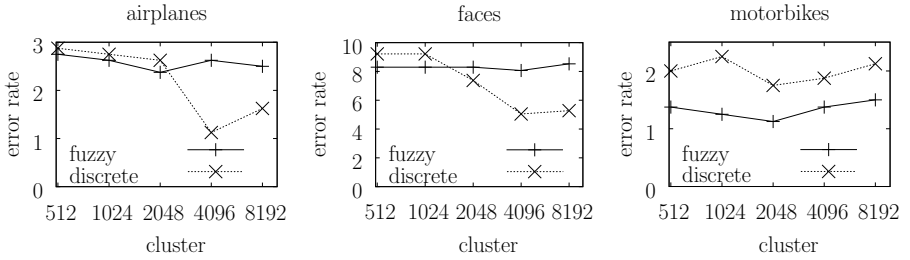

Fig. 5. Error rates for the Caltech tasks depending on the number of clusters using fuzzy histograms and discrete histograms.

fuzzy histograms are used. The results in Table 1 show that the use of fuzzy histograms does not yield a significant improvement over the baseline method.

The results from evaluating the proposed method for brightness normalization can be seen in Table 1. It can be seen that strong improvements are possible here. A significant improvement is observed in the faces task because some of the images were taken indoors and some images where taken outdoors.

Finally, SVMs were tested and the results cannot be improved. Thus, apart from losing the possibility to visually see which patches are discriminative for which class we loose classification performance using SVMs.

The result presented in [6] is much better for the faces task, because a specialized method for face detection was applied to the data.

6 Conclusion and Outlook

In this paper we extended a method for object classification in cluttered scenes into different directions. We proposed to use multi-scale feature, Sobel features, generalized histograms, and brightness normalization. We could show experimentally that multi-scale features and brightness normalization strongly improve the results, and that generalized histograms can be used to reduce computation time in training with only slight degradation in classification performance. Using Sobel features did not improve the results. It might be an interesting option to

apply PCA transformation to the gray values and Sobel features separately. Furthermore, we plan to explicitly model local variability in images. Another point where improvements are probably possible is to consider spatial information along with the extracted patches. All spatial information is currently discarded.

The results of the recent evaluation within the PASCAL Visual Object Classes Challenge[3] underline the quality of the approach.

Acknowledgments

We would like to thank Etienne Loupias for providing the source code for his salient point detector. This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under contract NE-572/6.

References

1. J. N. Darroch and D. Ratcliff. Generalized Iterative Scaling for Log-Linear Models. *Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.
2. T. Deselaers, D. Keysers, and H. Ney. Features for Image Retrieval – A Quantitative Comparison. In *DAGM Symposium, Pattern Recognition*, Tübingen, Germany, pages 228–236, Sep. 2004.
3. T. Deselaers, D. Keysers, and H. Ney. Discriminative Training for Object Recognition using Image Patches. In *CVPR*, San Diego, CA, in press, June 2005.
4. G. Dorko and C. Schmid. Selection of Scale-Invariant Parts for Object Class Recognition. In *ICCV*, volume 1, Nice, France, pages 634–640, Oct. 2003.
5. R. Fergus, P. Perona, and A. Zissermann. Object Class Recognition by Unsupervised Scale-Invariant Learning. In *CVPR*, Blacksburg, VG, pp. 264–271, June 2003.
6. M. Fussenegger, A. Opelt, A. Pinz, and P. Auer. Object Recognition Using Segmentation for Feature Detection. In *ICPR*, vol. 3, Cambridge, UK, pp. 41–48, Aug. 2004.
7. D. Keysers, F.-J. Och, and H. Ney. Maximum Entropy and Gaussian Models for Image Object Recognition. In *DAGM Symposium, Pattern Recognition*, Zürich, Switzerland, pages 498–506, Sep. 2002.
8. D. Keysers, C. Gollan, and H. Ney. Local Context in Non-linear Deformation Models for Handwritten Character Recognition. In *ICPR*, vol. 4, Cambridge, UK, pages 511–514, Aug. 2004.
9. T. Kölsch, D. Keysers, H. Ney, and R. Paredes. Enhancements for Local Feature Based Image Classification. In *ICPR*, vol. 1, pp. 248–251, Aug. 2004.
10. B. Leibe and B. Schiele. Scale Invariant Object Categorization Using a Scale-Adaptive Mean-Shift Search. In *DAGM Symposium, Pattern Recognition*, pp. 145–153, Aug. 2004.
11. E. Loupias, N. Sebe, S. Bres, and J. Jolion. Wavelet-based Salient Points for Image Retrieval. In *ICIP*, vol. 2, Vancouver, Canada, pages 518–521, Sep. 2000.
12. A. Martinez and A. Kak. PCA versus LDA. *IEEE TPAMI* 23(2):228–233, Feb. 2001.
13. A. Mohan, C. Papageorgiou, and T. Poggio. Example-based Object Detection in Images by Components. *IEEE TPAMI*, 23(4):349–361, Apr. 2001.
14. M. Weber, M. Welling, and P. Perona. Unsupervised Learning of Models for Recognition. In *ECCV*, vol. 1, Dublin, Ireland, pp. 18–32, June 2000.

[3] `http://www.pascal-network.org/challenges/VOC/`