Properties of Patch Based Approaches for the Recognition of Visual Object Classes

Alexandra Teynor[1], Esa Rahtu[2], Lokesh Setia[1], and Hans Burkhardt[1]

[1] University of Freiburg, Department of Computer Science
Georges Köhler Allee 052, 79110 Freiburg, Germany
{teynor, setia, Hans.Burkhardt}@informatik.uni-freiburg.de
[2] University of Oulu, Department of Electrical and Information Engineering,
PO Box 4500, 90014 Oulu, Finland
erahtu@ee.oulu.fi

Abstract. Patch based approaches have recently shown promising results for the recognition of visual object classes. This paper investigates the role of different properties of patches. In particular, we explore how size, location and nature of interest points influence recognition performance. Also, different feature types are evaluated. For our experiments we use three common databases at different levels of difficulty to make our statements more general. The insights given in the conclusion can serve as guidelines for developers of algorithms using image patches.

1 Introduction

The amount of digital documents increases daily, and with it the need to organize this torrent of data in order to retrieve something again. Especially for digital images no ideal solution has been found yet. The manual annotation of images is very labor intensive, so the vast majority of images will remain unannotated. Techniques for content based image retrieval (CBIR) are able to find similar images based on pixel content only, however, usually the definition of similarity is on a color and texture level, not on a semantic level. Most users do not want to find things with just the same texture and color, but want to find semantic entities, images with particular objects like cows, sheep or cars. This is why the main focus of research is now drawn to the recognition of visual object classes rather than the already widely researched area of traditional CBIR, as surveyed e.g. in [1].

1.1 Basic Principles

Currently, the most promising approaches for the recognition of visual object classes are based on the use of image patches. The advantages are easy to see: local representations can deal with variability in object shape and partial occlusions. The majority of these approaches follow an easy basic pattern: first, points or areas of high information content become identified in images. For this, so

K. Franke et al. (Eds.): DAGM 2006, LNCS 4174, pp. 284–293, 2006.

called interest point detectors or covariant region detectors are used. A survey about them can be found, e.g. in [2]. In the next step, features get extracted from these locations. Now models can be built for each class to be recognized or the feature vectors can be used directly. Depending on the model, different classifiers (e.g. SVMs [3], Winnows [4], Bayes [5]) can be used.

1.2 Related Work

A fair amount of work has already been done using image patches for classification and retrieval. One of the early approaches was by C. Schmid et al. [6]. She proposed calculating local gray value invariants at interest points for image retrieval. Weber et al. [7] and Fergus et al. [5] introduced a so called "constellation model", i.e. image patches in a probabilistic spatial arrangement, to decide whether a certain object is present in a scene or not. Agarwal et al. [4] classified and localized objects in an image using binary vectors coding the occurrences and spatial relations of patches. Leibe et al. showed in [8] a method to simultaneously categorize and segment objects using an implicit shape model. D. Lowe [9] proposed highly distinctive SIFT features in order to detect objects reliably in a scene. More recent work on this topic was conducted, e.g. by Deselaers et al. [10], who used histograms of patch cluster memberships in order to compare different classification methods. Opelt et al. [11] used a great variety of features and classifiers in a boosting framework to distinguish the best choice for each class.

The authors of the previously mentioned works had to decide at some point where to take patches, how many and at which size. Most of these decisions were done empirically in the course of the work, or were predetermined by the models chosen. E.g. the joint probability model used in the constellation model prohibits the use of much more than 7 parts. Only few works that we are aware of deal explicitly with these questions, e.g., Deselaers et al. [10] conducted some experiments for different patch sizes. The choice of the local descriptor type was investigated in [12] for matching and in [13] for object categorization. However, there the size of the patches is selected by automatic detectors, which does not necessarily mean that the size is optimal for object categorization.

In this work, we want to investigate how the factors number and size of image patches, descriptor type and nature of interest points influence retrieval quality.

1.3 Outline of the Paper

After the introductory section we briefly describe the types of local descriptors and the interest point detectors used. In section 3 we explain the test setting and experiments we performed. In section 4 we describe the results of our experiments and discuss them. In the last section we set out our conclusions and give recommendations for developers of patch based approaches for object categorization.

2 Methods and Algorithms

2.1 Interest Point and Affine Covariant Area Detection

In most of our experiments we use an interest point detector to find prominent places for the extracted patches. Some guidelines on the choice of the detector can be found from evaluation papers like [2,14,12], but these concentrate mainly on repeatability and information content of such points, which is not necessarily the key issue in patch based object categorization. For the majority of the experiments we selected the wavelet based interest point detector presented by Loupias and Sebe in [15]. This choice was motivated by good results in evaluation papers [14], image retrieval [16] and object categorization [10]. Of course other detectors, especially scale invariant ones could have been used, however we wanted to see the direct influence of patch size. An extensive region detector evaluation was out of scope for this paper.

In addition to the location, the ideal shape of the patch is also a question. Simple approaches use round or square patches centered at interest locations, more sophisticated solutions use affine region detectors. To test how they perform compared to each other, we selected two affine region detectors as examples: the Harris-Affine detector [17] and the maximally stable extremal regions (MSER) detector [18].

2.2 Feature Extraction

Once interest points or covariant areas have been found, features can be extracted. In the following, we briefly describe the features used in this evaluation. Some of them are subject to a PCA (principal component analysis) in order to get a more compact representation, details about this can be found in section 4.1

Gray values: The simplest way to get a description of the area around the interest point is to directly use the gray values in a window with side length $2d+1$ (d being the patch radius) centered around the interest point.

Multi-Scale Autoconvolution: The Multiscale Autoconvolution (MSA) is an $\mathbf{R}^2 \rightarrow \mathbf{R}^2$ mapping which is invariant with respect to affine transformations of the input function. This makes it possible to use MSA transform values as features for affine invariant classification. The basic idea behind MSA is to apply probabilistic approaches to the affine coordinate system. For an image function $f(x, y)$ the MSA transform is

$$If(\alpha, \beta) = E[f(\alpha(x_1 - x_0) + \beta(x_2 - x_0) + x_0)], \tag{1}$$

where α, $\beta \in \mathbf{R}$, E is the expected value and x_0, x_1, x_2 are random points with probability density given by $f(x, y)/||f(x, y)||_{L^1}$. A comprehensive introduction to MSA can be found in [19].

Haar integral based invariants: Schulz-Mirbach [20] introduced image features based on Haar integrals invariant to transformation groups. These are

constructed as follows: Let $\mathbf{M} = \mathbf{M}(i,j), 0 \leq i < N, 0 \leq j < M$ be an image, with $\mathbf{M}(i,j)$ representing the gray-value at the pixel coordinate (i,j). Let G be the transformation group of translations and rotations with elements $g \in G$ acting on the images, such that the transformed image is $g\mathbf{M}$. An invariant feature must satisfy $F(g\mathbf{M}) = F(\mathbf{M}), \forall g \in G$. Such invariant features can be constructed by integrating $f(g\mathbf{M})$ over the transformation group G

$$I(\mathbf{M}) = \frac{1}{|G|} \int_G f(g\mathbf{M}) dg$$

which for a discrete image is approximated using summations. By using k different kernel functions f we get a k-dimensional feature vector for each location.

Scale Invariant Feature Transform (SIFT): Scale invariant feature transform (SIFT) introduced by Lowe in [21] is based on histograms of Gaussian weighted gradient orientations around scale invariant interest points. To be more comparable, we did not use the SIFT built-in interest point detector, but the same locations and scales as for the other features.

3 Databases and Test Setting

For our evaluation we used 3 image databases at different levels of difficulty. We only used gray value information. The most simple database is the ETH80 database introduced in [22]. Here 10 different objects from 8 different object classes are photographed in front of a uniform background. For each object, 41 views are taken at different angles. For this database, the classifier had to decide which of the 8 object classes is present. Tests were performed in a leave-one-object-out approach.

The second image sets are from the Caltech dataset [1]. We chose to take the most commonly used collections "airplanes_side" (1074 images), "faces" (450 images) and "motorbikes_side" (826 images). For this database an object present/absent task has to be solved. As a counter class, a set of mixed "background" (900 images) images is used. The individual objects differ in appearance and location, but are about the same size and orientation. The background is cluttered. We divided each collection randomly into two halves, from which one was used for training and the other one for testing.

A clearly more difficult categorization task is present in the Graz02 database[2]. This database has four object categories: "cars" (420), "persons" (311 images), "bikes" (365 images) and a so-called "none" category (380 images) which was used as a counter class. In all the categories, objects suffer from severe occlusions and have a highly variable appearance and pose, reflecting real world scenes more accurately. Experiments performed with this database used the same setting introduced with the Caltech database. Some example images from the three databases can be seen in Figure 1.

[1] http://www.robots.ox.ac.uk/~vgg/data3.html
[2] http://www.emt.tugraz.at/~pinz/data/GRAZ_02

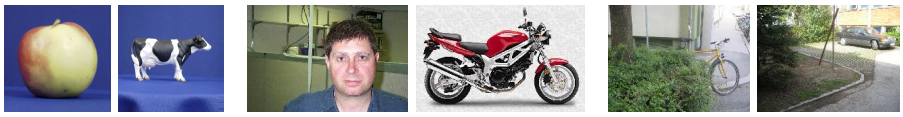

Fig. 1. Sample images: ETH80 (left), Caltech (middle), Graz02 (right)

3.1 Classification Procedure

In this paper our goal was to examine properties of patches and features extracted from them, so we wanted to keep the classification procedure simple. To do this, we apply a nearest neighbor classifier with a suitable distance measure. We first fit a multivariate Gaussian distribution to all feature vectors of each image, obtaining a mean vector μ and the full covariance matrix Σ. To determine the distance, we use the symmetric form of the Kullback-Leibler Divergence, for which a closed form expression can be derived:

$$KL[p_1(x)||p_2(x)] = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx + \int p_2(x) \log \frac{p_2(x)}{p_1(x)} dx \qquad (2)$$

4 Experimental Results and Discussion

In the following, we show the outcome of our experiments. Due to space constraints, in the result tables for the ETH80 database, the apple and horse class are not shown, since these objects are similar in appearance to tomato and pear or horse and dog respectively. All categories are contained in the "all" column.

4.1 Feature Types

One of the first questions is which type of features to select. Different feature types have different properties for different tasks. We tested various features already described in section 2 for their suitability for object class recognition. For this experiment, the radius of the patches was 20, we used the Loupias interest point detector and 100 points were selected per image. Interest points closer to the border than the radius were omitted. The gray value features were reduced to 20, the SIFT and MSA features to 10 dimensions via PCA, since an estimation of a multivariate Gaussian distribution with full dimension would be too imprecise. The dimensions were chosen by the amount of variance covered by the corresponding eigenvectors. The different dimensionality of the features is due to the difference in initial feature size. For the Haar integral features, we used 20 kernel functions. The results are summarized in table 1.

For the features tested, the SIFT features performed best for the ETH80 database and the Caltech database. The results are especially good if we think about the simplicity of our classifier. The gray value based features performed in the upper range for all three databases, making them suitable for systems in need of simple feature extraction methods. The MSA and Haar invariants did

Table 1. Classification rate of different feature types (in %)

ETH80	car	cow	cup	dog	pea	tom	all
Gray	86.6	49.5	75.9	70.2	93.2	90.7	73.9
MSA	71.0	70.0	65.6	52.7	86.1	90.7	68.4
Haar	81.7	66.3	64.3	62.4	89.0	67.6	67.0
SIFT	85.9	56.1	78.5	55.9	98.3	88.1	74.6

	Caltech database			Graz02 database		
	airp.	faces	mot.	bike	cars	pers.
Gray	91.1	93.0	90.7	72.7	68.8	79.5
MSA	85.8	91.1	91.9	72.7	67.3	73.1
Haar	90.0	91.0	90.6	71.3	63.0	70.8
SIFT	97.3	95.9	91.4	71.9	58.8	67.1

not perform as well for this task. We can also notice that the results for the ETH80 database are worse than the global approaches introduced in [22]. For nearly segmented, unoccluded objects, global methods work better.

4.2 Patch Size

Another important question is the patch size. If we select it too small, we are in danger of getting unspecific parts, if we select it too big, we might end up with patches that no longer have generalization capabilities. In this experiment, we use gray values as baseline features (PCA, 20D unless otherwise stated). When judging the results, we have to keep in mind that smaller patches do not lose as much information with PCA as larger patches, since the initial data size is much smaller. In the Caltech database, the motorbikes have the same size relative to the image, but the image sizes vary, so we scaled them to the same height of 250. The objects in Graz02 database are of very different size, so we omit this database for this experiment, since no single patch size makes sense here.

Table 2. Classification rate of different patch size (in %)

ETH80	car	cow	cup	dog	pear	tom	all
2 (10D)	93.9	40.7	78.3	46.1	85.9	63.2	64.2
5	91.5	40.7	75.4	63.9	80.2	44.4	63.7
10	86.1	49.5	78.3	77.8	88.8	72.9	71.5
20	86.6	49.5	75.9	70.2	93.2	90.7	73.9
30	85.9	51.7	74.4	63.9	97.3	94.6	74.1

Caltech	airp.	faces	mot.
2 (10D)	96.9	94.4	97.6
5	95.9	92.6	97.2
10	94.9	86.4	95.6
20	93.3	86.5	94.6
25	93.7	87.6	93.7

For the segmented objects in the ETH80 database, on average bigger patches perform better. Looking at the classification results for different objects reveals more details: for rather uniform objects with a smooth outline like pears or tomatoes, bigger patches clearly perform better. This is likely because a bigger part of the silhouette carries more information, the smaller the parts we have the more similar they are. For more detailed objects, small parts usually work better. For the Caltech database, smaller patches seem to work best in all cases, since we do not have smooth objects there. Figure 2 gives us an impression how a patch looks at the same interest point in different sizes.

Fig. 2. Example patch for a motorbike wheel with radius 5, 10 and 20

4.3 Number of Interest Points

The next question we address is the number of interest points. In this experiment, we take the N most salient points given by the Loupias detector. At these points the gray values are taken in a window with radius 20, the feature dimension is reduced to 20 via PCA.

Table 3. Classification rate for different numbers of interest points (in %)

									Caltech database			Graz02 database		
ETH80	car	cow	cup	dog	pear	tom	all		airp.	faces	mot.	bike	cars	pers.
20(10D)	63.7	40.7	69.3	51.0	87.1	92.2	63.1	20(10D)	87.7	88.3	85.8	61.1	61.3	71.1
50	83.4	40.0	72.0	68.5	88.8	92.9	71.4	50	92.7	93.0	89.3	74.0	67.0	74.0
100	86.6	49.5	75.9	70.2	93.2	90.7	73.9	100	91.0	93.0	90.9	72.7	68.8	79.5
200	88.3	52.4	73.2	65.6	94.6	89.8	74.4	200	90.6	91.1	92.0	73.7	66.3	76.9
500	88.8	52.4	74.9	60.7	95.9	91.7	75.0	500	89.8	90.1	91.9	74.3	67.8	74.9

When dealing with objects in front of a uniform background as in the ETH80 database, taking more interest points converges to an optimum for high numbers, since most of the patches convey object information. For databases with a (highly) cluttered background this is no longer the case. An intermediate range of about 100 interest points has shown to be sufficient, given our classification method and these databases. Taking too many Loupias interest points usually means taking more background clutter. Results are listed in table 3, in figure 3 we illustrate the area that is covered by 20, 50 and 200 interest points for a sample image.

Fig. 3. Area covered by N most salient points, N= 20, 50 and 200

4.4 Interest Points vs. Random Points

What is the role of the interest point detector in the selection of the patches? Does it give a clear advantage over taking random points? The following experiment should clarify this. We calculate the feature vectors (again for simplicity PCA reduced gray values, window radius 20, 20 dimensions) at a varying number of random points.

Table 4. Classification rate for different numbers of random points (in %)

ETH80	car	cow	cup	dog	pear	tom	all
50(10D)	24.6	29.5	52.9	26.3	52.0	85.4	45.2
100	24.1	27.1	58.3	34.4	25.1	74.9	43.8
200	51.0	34.6	72.4	42.7	62.4	79.5	57.4
500	70.7	49.5	75.6	45.9	80.0	91.0	67.3
1000	78.3	48.8	77.3	53.2	88.8	92.0	71.2

	Caltech database			Graz02 database		
	airp.	faces	mot.	bike	cars	pers.
50(10D)	88.5	87.9	82.0	65.7	58.0	69.7
100	87.8	89.6	86.2	67.6	57.5	73.7
200	91.0	91.9	87.8	68.9	57.5	68.8
500	92.4	90.5	85.1	70.2	66.8	74.3
1000	92.4	91.4	84.9	73.5	65.3	73.1

For our experiments, computing features at interest points is superior to random points, as can be seen in table 4. This is especially visible at the ETH80 and the Graz02 databases. Even for 1000 random points, the classification accuracy obtained with fewer interest points cannot be achieved. The exception to the rule is the airplanes category in the Caltech database. For this dataset, a uniform background (=sky), where no interest points are found, is a discriminative property. This confirms that context information can be beneficial for categorization. For the faces class, starting from 200 points, it does not make a difference whether to take interest or random points.

4.5 Shape of Interest Points - Fix vs. Affine Invariant

In our last experiment, we wanted to see whether it is beneficial to use features calculated from covariant regions instead of using windows of fixed geometry (squares or circles). A problem with fixed patches is that their content might change considerably when the viewing angle or the scale of an object changes, however, the automatically detected orientations and scales do not need to be ideal for categorization. We tested the affine harris detector and the MSER detector, together with two feature extraction methods, SIFT and MSA. As MSA is affine invariant, it can be directly applied to the patches. For SIFT features, the elliptical regions have to be normalized to circles. For the calculation, we used the binaries provided by C. Schmid and K. Mikolajczyk [3]. The number of interest points detected by these detectors varied a lot depending on the image. We used parameters so that around 100-400 patches were found. This number is slightly higher than in the case with fixed geometry, since many of the affine covariant areas were too small to cover the object adequately.

The final classification results for the Caltech and the Graz02 databases are shown in table 5, together with corresponding results for a fixed geometry. We had to omit the ETH80 database, because the region detectors were not able to find reasonable regions from all of the images. Some objects, like pears or tomatoes, seem to be too smooth for covariant detectors to converge. Especially for the SIFT features, the combination with the MSER detector seems to have a clear advantage over fixed patches. However, the classification performance did not improve in all cases. Especially using the harris affine detector degraded

[3] http://www.robots.ox.ac.uk/~vgg/research/affine/

Table 5. Classification rates for different affine covariant patch detectors in % (ha = harris affine, mr = MSER)

	Caltech			Graz02		
	airp.	faces	mot.	bike	cars	pers
MSA	85.8	91.1	91.9	72.7	67.3	73.1
MSA ha	79.0	84.6	89.9	69.4	54.3	67.3
MSA mr	75.9	75.1	87.1	69.2	62.8	59.8

	Caltech			Graz02		
	airp.	faces	mot.	bike	cars	pers
SIFT	97.3	95.9	91.4	71.9	58.8	67.1
SIFT ha	83.5	75.7	78.8	73.7	57.5	63.9
SIFT mr	95.0	97.0	96.9	74.5	60.8	62.7

the results. We assume that the stable invariant areas found are not necessarily optimal in a categorization sense.

5 Conclusions

In this paper we addressed some fundamental questions about the use of patches in the categorization of visual object classes. We could show that feature type, size, number, shape and location of patches does influence the retrieval performance, in some cases significantly. The selection of the feature type depends on the image class to be recognized. This confirms that an automatic selection procedure for features as introduced by Opelt et al. in [11] is beneficial in order to get optimal results.

For detailed objects, smaller patches usually work better, for smooth and uniform objects, bigger patches are necessary to cover object information. Interest point detectors are preferable over random selection to determine the location for patches, as good retrieval results can be achieved with relatively few patches, at least for our simple classifier. This is especially true for images with prominent objects or segmented images, and holds less for images with much background clutter. Only in extreme cases, random selection is superior, especially when homogeneous areas, where no interest points are found, are discriminative. An intermediate number of interest points (usually a few hundred) should be extracted from moderately cluttered images, taking too many or too few points spoils recognition performance here. For segmented images, taking more patches converges to some optimum, since no corruptive background patches spoil recognition accuracy.

Affine covariant methods provide an elegant way to choose the shape of a patch, increasing the performance on some occasions. An interesting research issue is to further investigate to what extent the automatically chosen areas are advantageous for object categorization.

Acknowledgment. This work has been funded by the German Federal Ministry of Education and Research, project I-Search, grant No. 01IRB02B and the Muscle NoE, contract No. 507752.

References

1. Santini, S., Gupta, A., Smeulders, A., Worring, M., Jain, R.: Content based image retrieval at the end of the early years. **22** (2000) 1349–1380
2. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. IJCV **37** (2000) 151–172
3. Wallraven, C., Caputo, B., Graf, A.: Recognition with local features: the kernel recipe. In: Proc. ICCV. (2003)
4. Agarwal, S., Awan, A., Roth, D.: Learning to detect objects in images via a sparse, part-based representation. IEEE TPAMI **26** (2004) 1475–1490
5. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proc. CVPR. Volume 2., Madison, WI (2003) 264–27
6. Schmid, C., Mohr, R.: Local greyvalue invariants for image retrieval. IEEE TPAMI **19** (1997) 530–535
7. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for recognition. In: Proc. ECCV, Dublin, Ireland (2000)
8. Leibe, B., Schiele, B.: Interleaved object categorization and segmentation. In: Proc. BMVC, Norwich, UK (2003)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
10. Deselaers, T., Keysers, D., Ney, H.: Discriminative training for object recognition using image patches. In: Proc. CVPR. Volume 2., San Diego, CA (2005) 157–162
11. Opelt, A., Pinz, A., Fussenegger, M., Auer, P.: Generic object recognition with boosting. IEEE TPAMI **28** (2006) 416–431
12. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE TPAMI **27** (2005) 1615–1630
13. Mikolajczyk, K., Leibe, B., Schiele, B.: Local features for object class recognition. In: Proc. ICCV. Volume 2. (2005) 1792–1799
14. Sebe, N., Lew, M.S.: Comparing salient point detectors. PR Letters **24** (2003) 89–96
15. Loupias, E., Sebe, N.: Wavelet based salient points for image retrieval. Technical report, Laboratoire Reconnaissance de Formes et Vision, INSA Lyon (1999)
16. Halawani, A., Burkhardt, H.: Image retrieval by local evaluation of nonlinear kernel functions around salient points. In: Proc. ICPR. (2004)
17. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. IJCV **60** (2004) 63–86
18. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proc. BMVC, Cardif, UK (2002)
19. Rahtu, E., Salo, M., Heikkilä, J.: Affine invariant pattern recognition using multiscale autoconvolution. IEEE TPAMI **27** (2005) 908–918
20. Schulz-Mirbach, H.: Anwendung von Invarianzprinzipien zur Merkmalgewinnung. PhD thesis, TU Hamburg-Harburg (1995) Reihe 10, Nr. 372, VDI-Verlag.
21. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proc. ICCV, Corfu, Greece (1999) 1150–1157
22. Leibe, B., Schiele, B.: Analyzing contour and appearance based methods for object categorization. In: Proc. CVPR, Madison, WI (2003)