Discriminative Training for Object Recognition Using Image Patches

Thomas Deselaers, Daniel Keysers, and Hermann Ney
Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University – 52056 Aachen, Germany
{deselaers,keysers,ney}@cs.rwth-aachen.de

Abstract

We present a method for automatically learning discriminative image patches for the recognition of given object classes. The approach applies discriminative training of log-linear models to image patch histograms. We show that it works well on three tasks and performs significantly better than other methods using the same features. For example, the method decides that patches containing an eye are most important for distinguishing face from background images. The recognition performance is very competitive with error rates presented in other publications. In particular, a new best error rate for the Caltech motorbikes data of 1.5% is achieved.

1. Introduction

An important open problem in computer vision is the learning and recognition of objects in cluttered scenes. Different approaches to address this problem have been described in the literature. One very promising approach assumes that the objects to be learned and recognized consist of a collection of parts. This assumption has some immediate advantages: Changes in the geometrical relation between image parts can be modeled to be flexible or even to be ignored and the algorithm can focus on those image parts that are most important to recognize the object. It is also evident that this approach can handle occlusions well. If parts of an object are occluded in an image, the remaining visible parts may still be used to recognize the object correctly and to learn about the appearance of this object from this instance.

When applying this paradigm of classification by image parts, we must take two decisions: At which points in the image do we extract image patches that should capture the object parts? Given the image patches, how do we decide which class of object is present in the image? Here, we use image patches that are extracted at points of interest and also at regularly spaced intervals. We use an available interest point detector [11], while we could also use the local variance or entropy [4, 13].

For the decision rule, we propose to use a discriminative model that takes as input the frequency of occurrence of the clusters that the patches are assigned to. In this work, we compare this discriminative log-linear model to other models that operate on the same features, including naive Bayes, maximum likelihood of the class conditional probability, and a nearest neighbor model. Furthermore, we compare the complete setup to a direct voting approach.

Related work includes Mohan and colleagues [12] who use predetermined parts of human bodies to detect humans in cluttered scenes. Dorko and Schmid [3] use image patches to classify cars, but the extracted patches from the training set are hand-labeled 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 only need weak supervision in training, i.e. only information about the presence of an object in the image. Fergus and colleagues [4] and Weber and colleagues [15] statistically model position, occurrence, and appearance of object parts.

2. Feature Extraction

Given an image, we use up to 1000 square image patches as features 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 onto 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. In informal experiments, this combination of points of interest and regular grid performed better than both methods alone. 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 corresponding closest cluster

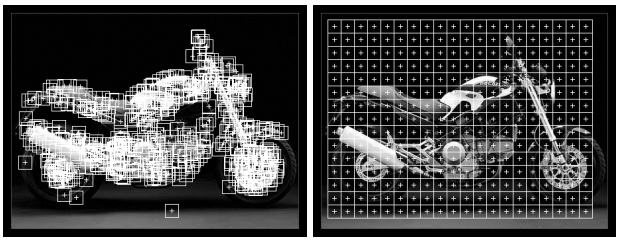

Figure 1: Patch extraction: salient points and uniform grid.

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 beloong 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.

Obviously, there exist alternatives to algorithmic choices made in the proposed method. For example, different interest point detectors can be used. However, experiments in other domains suggest that the choice of the interest point detector is not critical and often the local grey value variance or entropy is already a sufficient criterion, provided that enough image patches are extracted [13, 14]. Furthermore, the geometrical relation between the extracted patches is completely neglected in the approach presented here. While this relation could be used to improve classification accuracy, it remains difficult to achieve an effective reduction of the error rate in various situations by doing so.

3. Decision Rules

Having obtained the representation by (histograms of) image patches, we need to define a decision rule for the classification of images. In the following sections we briefly present different methods that use these representations. Note that most of the decision rules as they are presented here are simplified by the fact that in the experiments we assume a uniform prior distribution $p(k) = 1/K$.

3.1. Global Patch Search and Direct Voting

Global patch search and direct voting was proposed in [13] and uses the PCA-transformed image patches directly without computing a histogram representation. A KD-tree is created from the training image patches to admit efficient nearest neighbor searches. Using this KD-tree, each test image patch is assigned the class of its nearest neighbor using approximate search. The search is 'global', because all patches originating from one class are treated equally, independent of the image they were extracted from. The individual classifications of all patches are then combined by direct voting. The classification output is the class that most of the image patches have been assigned to. This method is known to obtain very competitive results in various tasks like face recognition, radiograph recognition, and character recognition [14] and therefore serves as a good baseline. Kölsch and colleagues [8] give a formalized description of the classification process and describe improvements that can be obtained by e.g. multi-scale patch extraction, a modified voting scheme, or invariant distance measures in the nearest neighbor search. In this work, we use the basic classification method as described above for comparison.

3.2. Nearest Neighbor

Using the histograms of image patches as a representation for the images, we can employ a simple nearest neighbor classifier. Usually, the nearest neighbor is a useful benchmark because it is a simple classifier with good performance in many applications. Here, we choose the Jensen-Shannon divergence to compare two histograms. This choice is based on findings in previous experiments [2], where this measure provided good performance across different tasks. The resulting decision rule for the nearest neighbor is then

$$X \mapsto r(X) = \arg\min_k \left\{ \min_{n=1\dots N_k} d(h(X), h(X_n)) \right\},$$

where $d(h, h') = \sum_{c=1}^{C} h_c \log \frac{2h_c}{h_c + h'_c} + h'_c \log \frac{2h'_c}{h'_c + h_c}$.

3.3. Naive Bayes

In the following approaches we use Bayes' decision rule

$$\begin{aligned} r(X) &= \arg\max_k \{p(k|X)\} \\ &= \arg\max_k \{p(k)\, p(X|k)\} \\ &= \arg\max_k \{p(X|k)\}, \end{aligned}$$

where the last equality holds due to $p(k) = 1/K$. Because we use the histogram representation of the images we let $p(k|X) := p(k|h(X))$ and $p(X|k) := p(h(X)|k)$.

In the naive Bayes approach, the assumption is made that the distributions of the feature vector components are conditionally independent. Thus, for the patch representation we assume that $p(X|k) = \prod_{l=1}^{L_X} p(x_l|k)$. As we assume uniform priors, the decision is not changed when we use the product of posterior probabilities. Furthermore, we apply

the logarithm to convert the product into a sum:

$$r(X) = \arg\max_k \left\{ \prod_{l=1}^{L_X} p(x_l|k) \right\} = \arg\max_k \left\{ \prod_{l=1}^{L_X} p(k|x_l) \right\}$$
$$= \arg\max_k \left\{ \sum_{l=1}^{L_X} \log p(k|x_l) \right\}$$
$$= \arg\max_k \left\{ \sum_{c=1}^{C} h_c(X) \log p(k|c) \right\},$$

Where we assume that these patch posterior probabilities are equal for patches within the same cluster: $p(k|x) = p(k|c(x))$. Finally, the cluster posterior probabilities are estimated from the relative frequencies on the training data:

$$p(k|c) = \frac{\sum_{n=1}^{N_k} h_c(X_{kn})}{\sum_{n=1}^{N} h_c(X_n)}$$

3.4. Generative Single Gaussian

Another baseline classification method is to use a single Gaussian density for the class-conditional probability $p(h|k) = \mathcal{N}(h|\mu_k, \Sigma)$ for each object class with pooled diagonal covariance matrices Σ. The parameters of the model are then estimated by the maximum likelihood method during training, maximizing $\prod_{k=1}^{K} \prod_{n=1}^{N_k} p(X_{kn}|k)$. In classification, Bayes' decision rule is used.

3.5. Discriminative Training

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

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

where $Z(h) = \sum_{k=1}^{K} \exp(\alpha_k + \sum_{c=1}^{C} \lambda_{kc} h_c)$ is the renormalization factor. (Note that also the generative Gaussian model can be rewritten in this form and we can furthermore always find a generative model that results in the same posterior distribution [6].) The maximizing distribution is unique and the resulting model is also the model of highest entropy with fixed marginal distributions of the features [6]. 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.

Figure 2: Examples from the Caltech data (airplanes, faces, motorbikes, background) and the medical radiographs.

3.6. Relation Between the Models

There exists a strong relation between the structure of the decision rule resulting from the naive Bayes, the Gaussian, and the log-linear model. In all three cases the decision rule can be rewritten as an $\arg\max$ operation of a linear function of the histogram representation:

$$r(X) = \arg\max_k \Big\{ \alpha_k + \sum_{c=1}^{C} \lambda_{kc} h_c(X) \Big\}$$

For the naive Bayes model we have $\alpha_k = 0$ and $\lambda_{kc} = \log p(k|c)$, for the Gaussian model the parameters $\{\alpha_k, \lambda_{kc}\}$ are a function of the parameters $\{\mu_k, \Sigma\}$, and for the log-linear model the parameters are trained directly.

In this formulation of the decision rule, evidently, patches assigned to those clusters c that have the highest absolute difference of coefficients $|\lambda_{kc} - \lambda_{k'c}|$ contribute the most to the discrimination between the classes k and k' according to the model. This correspondence is used to visualize the most discriminative patches in Section 5, where the sign of the difference $\lambda_{kc} - \lambda_{k'c}$ determines if the patch cluster contains indicators for class k or k'.

4. Database

Fergus and colleagues [4] 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 at http://www.robots.ox.ac.uk/∼vgg/data, 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 faces task consist 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 2.

To observe the performance of our method on a task with more than two classes, we also performed a key experiment using an in-house set of medical radiographs consisting of 2832 training images and 1016 test images. An example image is shown in Figure 2. This task consists of 24 classes which are very unevenly distributed. The data are courtesy of the IRMA project [9].

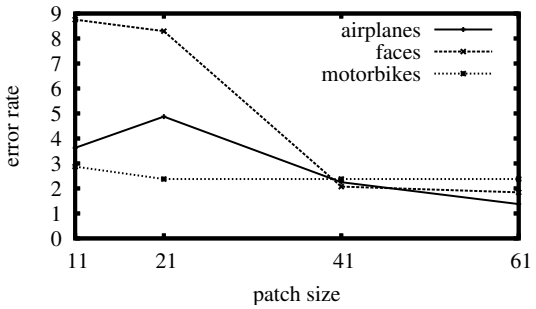

Figure 3: Discriminative model on the unscaled data: effect of patch size on the error rate.

5. Experimental Results

We evaluate two series of experiments for each of the Caltech tasks: In the first series, each image retains its original size and in the second series each image is scaled to the height of 225 pixels, the mean height of the input images. For each of the tasks, first the image patches are extracted. The PCA is estimated using the training data patches only and all patches are processed using the PCA coefficients. Then, the image patches from the training data are clustered to create the patch histograms for training and test data.

One parameter that must be determined for the experiments is the image patch size. We first use the images of original size, extract the features as described above, and apply the classification methods to these data. Figure 3 shows the error rate for these experiments using the discriminative model. The other classifiers behave similarly but yield larger error rates. It can be observed that the largest patch size (61×61) performs best. The resulting error rates for this patch size are shown in Table 1. They show that the discriminative model outperforms the other methods and are also very competitive with error rates presented in the literature for the same tasks as shown in Table 2. The second best approach is the naive Bayes model.

Visualizing the patches that are most discriminative according to the difference in coefficients from the discriminative approach shows an interesting effect. This effect results from the property that the images of the background class are generally smaller than the images from the other classes. The four most discriminative patches for the background and the motorbikes class are shown in Figure 4. It

Table 1: Error rates, size 61×61, original data, 512 clusters.

method	airplanes	faces	motorbikes
Global Patch Search	7.8	18.4	15.8
Nearest Neighbor	6.1	6.2	9.6
Naive Bayes	4.6	5.8	6.9
Generative Gaussian	15.4	30.0	19.0
Discriminative Model	**1.4**	**1.8**	**2.4**

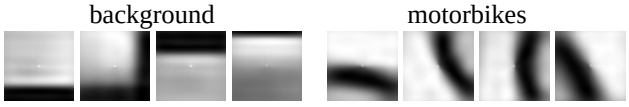

Figure 4: Most discriminative patches for size 61×61; background (left) vs. motorbikes (right).

can be clearly observed that the patches for the background class show image borders, while the patches for the motorbike class show parts of the wheels. On the one hand this shows that visually meaningful patches are learned to be discriminative for motorbikes. On the other hand, the significant difference in size is also learned by assigning more importance to patches that contain image borders and corners. This explains why enlarging the patches improves performance: The larger the patches are, the more of the image border is contained in the patches and thus for the smaller background images the most discriminative patches are those showing large amounts of image border.

Although we may state that the algorithm in fact learns to effectively discriminate between background and foreground images, this is not the result we are trying to obtain. While we believe that the error rates are still valid results, we are interested in the performance of the algorithm if it cannot exploit the difference in size of the image classes.

To avoid the effect of learning the borders of background images, we scale all images to the common height of 225 pixels, approximately the mean height across the data. Repeating our experiments to determine the best patch size we now obtain the error rates shown in Figure 5 for the discriminative model. Now larger patch sizes no longer perform better. The error rates for the smallest evaluated patch size of 11×11 are presented in Table 2 and compared to those from other publications and the best error rates from the first experiment. Performing further experiments with more cluster centers (thus using histograms with more bins) we observe that the error rate improves for the discriminative approach. The other methods, especially the generative Gaussian approach, improve only slightly if at all.

Again, the discriminative approach performs best among the investigated methods and gives competitive results. Especially the error rate of 1.5% for the motorbike task is the lowest published error rate we are aware of. The second best method now differs from task to task. (Note that the other publications give ROC equal error rates. The error rates presented here do not involve any adjustment of a threshold but still are very close to this concept: the misclassifications within the two classes are 14:16 for airplanes, 13:18 for faces, and 7:13 for motorbikes.)

In Figure 6 the top four discriminative patches are shown for each of the three tasks. We can observe that the patches for the foreground allow a meaningful visual interpretation in most of the cases: The airplane images contain more horizontal structures than the background images such that

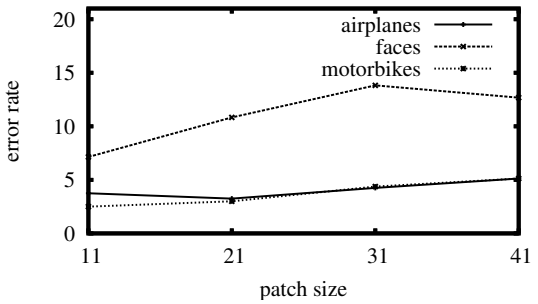

Figure 5: Discriminative model on the scaled data: effect of patch size on the error rate.

Table 2: Error rates on scaled Caltech data with 512/4096 clusters in comparison to results from other publications.

method	airp.	faces	mot.
512 clusters:			
Patch Search	4.8	8.5	21.5
Nearest Neighbor	9.4	18.7	5.5
Naive Bayes	8.5	17.1	9.9
Generative Gaussian	5.8	17.5	7.6
Discriminative Model	**3.8**	**7.1**	**2.5**
4096 clusters:			
Nearest Neighbor	11.6	19.9	14.5
Naive Bayes	5.6	11.3	7.5
Generative Gaussian	37.4	48.2	49.9
Discriminative Model	**2.6**	**5.8**	**1.5**
Statistical Model [4]	9.8	3.6	7.5
Texture features [2]	**0.8**	1.6	**7.4**
Segmentation [5]	2.2	**0.1**	10.4
Discr. Model (Table 1)	1.4	1.8	2.4

patches containing strong horizontal gradients are chosen to receive large weights. The first patch of the face class shows a patch that resembles an eye. This observation becomes clearer if we look at some patches from the training data that are assigned to this cluster as shown in Figure 7. Clearly, the algorithm has automatically learned that the eye is the visually most important feature to distinguish faces from background images. The second face patch can be interpreted as a part of the hair/forehead line while the third and fourth are not easily interpreted. For the motorbike task, all four patches show diagonal wheel/rim structures, which typically do not occur in background images. The most discriminative background patches change for the three tasks, which is due to different training images and to the discriminative training: For example, the first two background patches in the faces task are strong indicators for background versus faces, but this would not be true in the airplanes task, because here vertical structures are indicators for airplanes. (Note that the bright centered dot in

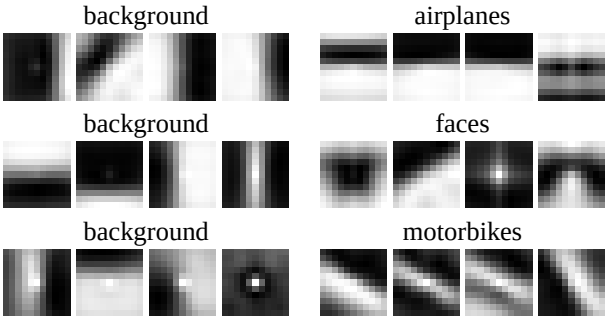

Figure 6: Most discriminative patches for the Caltech data (background and object class).

Figure 7: Cluster most discriminative for faces.

some of the patches is due to the PCA reconstruction and the mean image computed from patches supplied by the interest point detector, favoring images with strong gradients.)

Figure 8 shows typical examples for each of the three tasks with those positions marked at which highly discriminative features are extracted. We can observe that strong foreground indicators are located at horizontal structures for the airplanes task, at the eyes, hair/forehead, and clothes for the faces task, and at the wheel/rim and other diagonal structures for the motorbikes task. For the incorrectly classified images it can be observed that in the airplane image many vertical structures are found, that the face is too dark in comparison to the background, and that in the motorbike image a large amount of background features are present.

Table 3 shows the results using the above settings on the medical radiograph data. The error rate of 23% compares well to an error rate of 18% when using the image distortion model [7] which is a method that is known to produce excellent results on this corpus. Not that the obtained error rate of 23% was obtained without any adaptation of parameters and only serves to show that the approach can also be used for tasks with more than one class.

The experiments show that the presented approach works well for the data presented, where the foreground object forms a significant portion of the input image. It may be argued that it will be problematic for the approach to deal with cases where this is not the case. This (so far hypothetical) effect might be alleviated by using a significantly larger amount of training data. Furthermore, to our knowl-

Table 3: Error rates for the medical radiographs.

method	ER
Discriminative Model	23%
Image Distortion Model [7]	18%

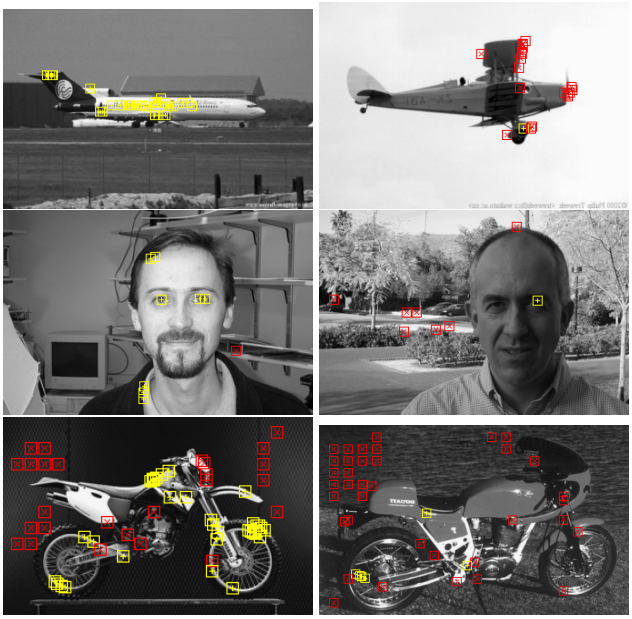

Figure 8: Typical examples of correct (left) and incorrect (right) classfications with positions of most discriminative patches for object (yellow) and background class (red).

edge this problem will occur for all generic learning and recognition approaches with the possible exception of those approaches that are tuned toward a specific application like face detection.

6. Conclusion

We presented a method for object classification that uses image patches and fully automatically learns which patches are discriminative for the given object classes. We compared the method to other methods using the same features. The obtained recognition performance compares favorably to those reported in other publications, in particular we observe a 1.5% error rate on the motorbikes task.

In the first series of experiments the discriminative training learns that the size of the image is a very discriminative feature for the classification by assigning a large weight to border and corner patches, which is not intended. To avoid this effect we scale the images to a common height in the second series of experiments. From the resulting clusters it can be observed that visually meaningful parts of the objects are learned, e.g. for faces the eyes and for motorbikes extracts from the wheels are most discriminative.

In future experiments, the next promising steps will be to use patches of different sizes to account for object parts at different scales and to allow the patches to distribute their votes to more than one bin of the histogram.

Acknowledgments

We would like to thank Etienne Loupias for providing the source code for his salient point detector, Javier Cano for his KD-tree library, and Andre Hegerath for help with the implementation. We would also like to thank Bernt Schiele for his proposition to use more cluster centers.

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*, vol. 43, no. 5, pp. 1470–1480, 1972.

[2] T. Deselaers, D. Keysers, and H. Ney, "Features for Image Retrieval – A Quantitative Comparison," *DAGM 2004*, pp. 228–236, Tübingen, Germany, Sept. 2004.

[3] G. Dorko, C. Schmid, "Selection of Scale-Invariant Parts for Object Class Recognition," *ICCV 2003*, pp. 634–640, 2003.

[4] R. Fergus, P. Perona, and A. Zissermann, "Object Class Recognition by Unsupervised Scale-Invariant Learning," *CVPR 2003*, pp. 264–271, Blacksburg, VG, June 2003.

[5] M. Fussenegger, A. Opelt, A. Pinz, and P. Auer, "Object Recognition Using Segmentation for Feature Detection," *ICPR 2004*, Cambridge, UK, Aug. 2004.

[6] D. Keysers, F.J. Och, and H. Ney, "Maximum Entropy and Gaussian Models for Image Object Recognition," *DAGM 2002*, pp. 498–506, Zürich, Switzerland, Sept. 2002.

[7] D. Keysers, C. Gollan, and H. Ney, "Local Context in Non-linear Deformation Models for Handwritten Character Recognition," *ICPR 2004*, vol. 4, pp. 511–514, Cambridge, UK, Aug. 2004.

[8] T. Kölsch, D. Keysers, H. Ney, and R. Paredes, "Enhancements for Local Feature Based Image Classification," *ICPR 2004*, vol. 1, pp. 248–251, Aug. 2004.

[9] T.M. Lehmann, M.O. Güld, C. Thies, B. Fischer, K. Spitzer, D. Keysers, H. Ney, M. Kohnen, H. Schubert, B.B. Wein, "Content-based image retrieval in medical applications," *Methods of Inf. in Medicine* 43(4): 354-361, April 2004

[10] B. Leibe and B. Schiele, "Scale Invariant Object Categorization Using a Scale-Adaptive Mean-Shift Search," *DAGM 2004*, pp. 145–153, Aug. 2004.

[11] E. Loupias, N. Sebe, S. Bres, and J. Jolion, "Wavelet-based Salient Points for Image Retrieval," *ICIP 2000*, vol. 2, pp. 518–521, Vancouver, Canada, Sept. 2000.

[12] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based Object Detection in Images by Components," *IEEE TPAMI*, vol. 23, no. 4, pp. 349–361, April 2001.

[13] R. Paredes, J. Perez-Cortes, A. Juan, and E. Vidal., "Local Representations and a Direct Voting Scheme for Face Recognition," *Workshop Pattern Recognition in Information Systems*, pp. 71–79, Setúbal, Portugal, July 2001.

[14] R. Paredes, D. Keysers, T.M. Lehmann, B.B. Wein, H. Ney, and E. Vidal, "Classification of Medical Images using Local Representations," *Bildverarbeitung für die Medizin*, pp. 171–174, Leipzig, Germany, March 2002.

[15] M. Weber, M. Welling, and P. Perona, "Unsupervised Learning of Models for Recognition," *ECCV 2000*, vol. 1, pp. 18–32, Dublin, Ireland, June 2000.