Enhancements for Local Feature Based Image Classification*

Tobias Kölsch, Daniel Keysers, Hermann Ney
Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{koelsch, keysers, ney}@informatik.rwth-aachen.de

Roberto Paredes
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Valencia, Spain – rparedes@dsic.upv.es

Abstract

Using local features with nearest neighbor search and direct voting obtains excellent results for various image classification tasks. In this work we decompose the method into its basic steps which are investigated in detail. Different feature extraction techniques, distance measures, and probability models are proposed and evaluated. We show that improvements are possible for each of the investigated enhancements. This shows that the important aspect of the framework is the decomposition of the training images into sets of local features for each class.

1. Introduction

For the task of appearance based image recognition, local and global approaches have been proposed. An inherent problem with global approaches is usually their susceptibility to local and global variations, e.g. slight changes of the viewpoint or illumination. Some of the methods that were proposed to circumvent this problem include the use of a distance measure that is invariant towards small global transformations [3, 10] or the use of representations that are invariant with respect to several transformations [1].

The use of local representations on the other hand provides the possibility to cope with local and global variability, e.g. translations and, to a certain extent, changes of the viewpoint. However, if the geometrical relationship between the features is to be taken into account, it has to be modeled explicitly. In [4] a map of the positions of the local features from two images that are compared is calculated using non-linear deformation models and in [2] the relative position, the scale, and the appearance of local features are modeled with a mixture density.

Many of the proposed local feature approaches compare the local features on a per image basis and model the global relation between the feature positions. In the approach presented here, the relation between local features is ignored and the comparison is not done on a per image basis. Instead, every local feature of the test image is compared to all features of all training images; therefore only the local image context is relevant to the matching and the final decision. This approach was introduced in [8] and obtained very good results on different tasks, e.g. for face verification [7]. The method is illustrated in Figure 1. To gain a deeper understanding of the approach, the feature extraction, the distance measure, and the probability model are investigated in more detail in this work.

*This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under contract NE-572/6.

2. Local Features for Image Classification

The local feature based method with direct voting was initially proposed in [8] and is used as a baseline method for our investigations. Here we give a short overview of this method followed by the proposed enhancements.

Local features are square sub images with a size of $I \times I$ pixels from an image. They are extracted at positions with a high local variance of the grey values. This is done in practice by defining a threshold $t \in \mathbb{R}$ and using only local features with a local variance higher than t. These are the extracts of the image that are expected to be well suited for discrimination. This approach has been shown to improve recognition results compared to taking all local features [6].

In the training phase, the features are extracted and a *principal components analysis* (PCA) is applied. This gives a representation of the features where the components are sorted by importance. The dimensionality of the representation is reduced by discarding all components with a higher index than D. These reduced features are then labeled with their class name and stored in a kd-tree to allow for a fast nearest neighbor search.

In testing, the $I \times I$ dimensional windows are also extracted at positions of high local variance, and their dimensionality is reduced using the PCA transform that has been computed in training. Then, for each local feature of the test image, the feature from the training images that is most similar with respect to the Euclidean distance is searched within the kd-tree. The nearest neighbors of the test features are grouped according to their class labels and the class which

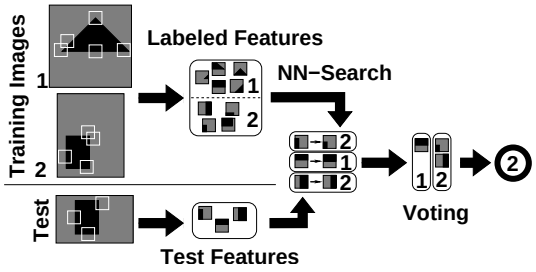

Figure 1. Schematic view of classification.

contains most of nearest neighbors is chosen as the classification result. This decision rule is usually referred to as *direct voting* as every local feature gives an equally weighted vote to the class it belongs to. A probabilistic interpretation of this approach is presented in Section 2.3.

2.1. Multi Scale Feature Extraction

In the baseline method, the extracted features all have the same size. However, by restricting the comparison to sub images of only one size, we implicitly assume that the objects are represented at the same scale in all images that are to be classified. Motivated by the results presented in [2], we relax this constraint by extracting features at different scales. To do so, a minimal feature size $I_{\min}$, a step size s, and a maximal feature size $I_{\max}$ are chosen. Then the extraction is performed at all pixel positions for windows with the width of $I_{\min} + s \cdot r < I_{\max}$ with $r \in \mathbb{N}$ and the extracts with low local variance are discarded. $I_{\max}$ is set to the size of the smaller image dimension. In the next step, the features are scaled to a fixed size $\hat{I}$. Then the features are PCA transformed and stored in the kd-tree. This extraction is done for the training and the test images, resulting in a potentially larger number of local sub images. However, this number of sub images can be adjusted by increasing the variance threshold t.

2.2. Tangent Distance

The similarity between two local features is measured using the Euclidean distance in the baseline method. However this measure is inherently sensitive to all kinds of transformations, as e.g. rotation, scaling, brightness changes, etc. The *tangent distance* (TD) [3, 10] gives invariance with respect to small global transformations that are known *a priori*. The transformations modeled are usually the 6 projective transformations eventually complemented by some problem specific transformations. To our knowledge the TD has only been used to compare entire images, so far. Here, it is used to measure the similarity of local features.

Let x be an image extract and $\tilde{x}$ a $I^2 \times L$ matrix containing its L used tangent vectors. The TD is then given by

$$\mathrm{d}(x,\, x') = \min_{\alpha \in \mathbb{R}^L} \{\|(x + \tilde{x} \cdot \alpha) - x'\|\}$$

Note that in this work the tangents are applied to the local feature vectors of the test image.

We use the TD in combination with the local features for the nearest neighbor search. However, the local features are PCA-transformed image extracts. The TD cannot be easily computed directly on the PCA-transformed vectors, as it is originally designed for images. However, the linearity of the PCA transformation allows us to transform the tangent vectors using the same PCA transformation matrix $M \in \mathbb{R}^{D \times I^2}$ as is used for the images and then calculate the TD in the reduced space:

$$M \cdot (x + \tilde{x} \cdot \alpha) = M \cdot x + (M \cdot \tilde{x}) \cdot \alpha$$

2.3. Probability Model

From the distance function, we compute the local feature posterior probability as

$$p(k|x) = \frac{\exp[-\frac{1}{2(\lambda\sigma)^2}\,\mathrm{d}(x, \hat{x}_k)]}{\sum_{k'=1}^{K} \exp[-\frac{1}{2(\lambda\sigma)^2}\,\mathrm{d}(x, \hat{x}_{k'})]}, \qquad (1)$$

where σ^2 denotes the empirical variance over the training features and λ is an empirical parameter. We denote by x the test feature vector and $\hat{x}_k$ its nearest neighbor from the class k according to the distance measure d. This approach is derived from the maximum approximation to the kernel density or Parzen window estimator. The baseline method uses the Euclidean distance, while we propose to use the TD instead. Furthermore, in the baseline method, a binary probability model is used. $p(k|x)$ is 1 if the nearest neighbor of x is from class k and 0 otherwise, which amounts to direct voting. This is a limiting case of the model (1) above for $\lambda \to 0$. The probability that the image X with the local features $x_1, \ldots, x_{N_X}$ is from class k is then computed using the *sum rule*, which is known to be well suited for noisy data [5]:

$$p(k|X) = \frac{1}{N_X} \sum_{n=1}^{N_X} p(k|x_n)$$

3. Databases

The extensions to the baseline method are evaluated on three different classification databases:

(a) *Image Retrieval in Medical Applications* (IRMA) is a corpus used in a cooperation of three departments of the RWTH Aachen University containing 1617 images of radiographies. The images are subdivided into 6

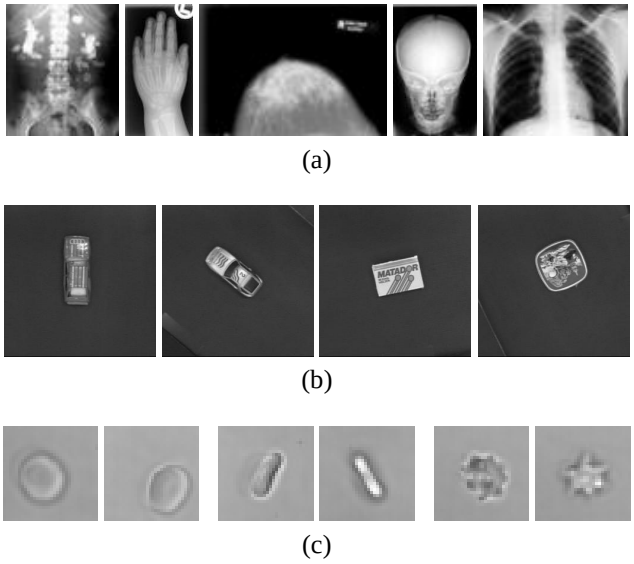

Figure 2. Example images from (a) IRMA, (b) Erlangen, and (c) Bloodcells.

classes depending on the body region they represent. The image size is variable. We scale all images such that the shortest side is 32 pixels and the aspect ratio is kept. Also a 2-bin histogram normalization is performed to increase the contrast of the images. Some example images can be seen in Figure 2(a). The best error rate on this database is 5.3% [4].

(b) *Erlangen* is an object recognition corpus from the Chair for Pattern Recognition of the University Erlangen-Nürnberg. It contains 6 tasks. Here the only task that is considered is object recognition with partial occlusion and changing illumination, as it is the most difficult one. The corpus contains 5 objects that are rotated in steps of $5°$. The images have a size of 256×256 pixels and are scaled down to 128×128. The best error rate on this corpus of 4.8% is reported in [9].

(c) *Bloodcells* is a corpus for red blood cell classification. It contains 5062 images of 3 types of red blood cells. The images are 64×64 pixels and are scaled down to 32×32 pixels. Histogram normalization is performed to improve the image contrast. It is quite a difficult classification task as can be seen from the best reported error rate of 15.3% and the human error of approximately 20% [1].

4. Experimental Results

Kernel densities are compared to direct voting on several corpora. Their use improves the recognition result in all cases. On IRMA, direct voting leads to 10.3% error and the kernel densities approach leads to 9.7%. On Erlangen direct voting results in 1.2% against 0.6% error for kernel densities and the result on Bloodcells is 17.7% without and 17.2% with kernel densities. In the following, we show that the use of kernel densities improves the results when combined with multi-scale feature extraction and with the TD. This suggests that the use of the distance to estimate the probability is generally better than voting directly.

The use of multi-scale features is compared to the extraction of features at only one scale. With exception of this, all parameters are maintained the same. The tests are performed on the IRMA corpus, once using direct voting and once with kernel densities. The results improve from 10.3% to 10.0% error with direct voting and from 9.7% to 9.4% for the kernel densities approach. This shows that the multi-scale extraction leads to improvements independent of the probability model.

The TD is compared to the Euclidean distance on IRMA. The transformations that are approximated by tangents are the affine trasnformations of the image plane and additive image brightness. With direct voting the result is improved from 10.3% to 7.7% by using the TD and if the kernel densities are used as a probability model the result is improved from 9.7% to 7.4% error. We observe a significant decrease of recognition error in both conditions.

We get another interesting result if we use the horizontal and vertical Sobel filter on the local features. With the Euclidean distance, the error rate improves from 10.3% to 8.9%. However, if the Sobel filter is used in combination with the TD, a result of 7.8% is obtained instead of the 7.7% that are obtained without Sobel filter.

On Bloodcells the TD is tested with the same tangents as for IRMA and additionally with a tangent for line thickness. This tangent was included, because the width of the borders of the cells varies strongly even within one class. The result is 17.2% error for the Euclidean distance and 13.5% error for the TD. Direct voting is used as a probability model. This result represents the best known outcome on this database.

The feature dimensionality reduction is typically done using PCA. We also have experimented with the *discrete cosine transform* (DCT) for dimensionality-reduction. This is done by transforming a feature into a wave image using the DCT and discarding the higher frequency components of the representation. Doing so, we obtain an error rate of 11.0% on IRMA compared to 10.3% with the PCA. However the advantage of the DCT is that it does not have to be computed on the training data and thus saves one processing step, while it leads to only slight degradation in performance.

The results on the three databases, along with the best other results published, are presented in the Tables 1 to 3. The approaches presented here lead to the best results on

Table 1. Results for the IRMA task.

Method	Error (%)
Euclidean distance, 1-NN [4]	15.8
Direct voting	10.3
Kernel density	9.7
Multi scale, Kernel density	9.4
Sobel, Direct voting	8.9
TD, Kernel density	**7.4**
Best other [4]	5.3

Table 2. Results for the Erlangen task.

Method	Error (%)
Direct voting	1.2
Kernel density	**0.6**
Best other [9]	4.8

Table 3. Results for the Bloodcell task.

Method	Error (%)
Euclidean distance, 1-NN	24.4
Direct voting	17.7
Kernel density	17.2
TD, Direct voting	**13.5**
Best other [1]	15.3

Bloodcells and on Erlangen. The results on Erlangen clearly show that the local feature approach is well suited to cope with occlusion.

5. Conclusion

In this paper we investigated enhancements for local feature based image classification. We showed that recognition is improved when using multi-scale features for the IRMA database. Using kernel densities instead of direct voting also improves recognition on all three used databases. Finally, applying the TD instead of the Euclidean distance leads to improvements as well. If the DCT is used instead of the PCA for feature dimensionality reduction the result deteriorates slightly, but the PCA estimation step on the training data can be saved. The experiments show that the local feature based approach is well suited to cope with partial occlusion. Observing that in each of the components feature extraction, distance measure, and probability model improvements of the baseline method are possible, we may assume that the main point of the method (that already makes the baseline method very powerful) is the following: in the search for similar image parts, all local features of one class are hypothesized at the same time and not on a per image basis, neglecting the position of the extracted local features.

We believe that adding global constraints to the local features approach could lead to improvements. A possible approach for this could be an image distortion model similar to that proposed in reference [4]. Also, combining the appearance based local features with other kinds of features, such as texture characteristics, could improve recognition. Finally, estimating the importance of features could also lead to improvements in recognition.

References

[1] J. Dahmen, J. Hektor, R. Perrey, and H. Ney. Automatic classification of red blood cells using gaussian mixture densities. In *Bildverarbeitung für die Medizin*, pp. 331–335, Munich, Germany, Mar. 2000.

[2] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Int. Conf. Computer Vision and Pattern Recognition*, volume 2, pp. 264–275, Madison, Wisconsin, USA, June 2003.

[3] D. Keysers, J. Dahmen, T. Theiner, and H. Ney. Experiments with an extended tangent distance. In *15th Int. Conf. on Pattern Recognition*, volume 2, pp. 38–42, Barcelona, Spain, Sept. 2000.

[4] D. Keysers, C. Gollan, and H. Ney. Classification of medical images using non-linear distortion models. In *Bildverarbeitung für die Medizin*, Berlin, Germany, pp. 366–370, Mar. 2004.

[5] J. Kittler, M. Hatef, R. P. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, Mar. 1998.

[6] T. Kölsch. Local features for image classification. Diploma thesis, Lehrstuhl für Informatik VI, RWTH Aachen, Aachen, Germany, Nov. 2003.

[7] K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, S. Srisuk, M. Petrou, W. Kurutach, A. Kadyrov, R. Paredes, B. Kepenekci, F. B. Tek, G. B. Akar, F. Deravi, and N. Mavity. Face verification competition on the XM2VTS database. In *4th Int. Conf. Audio and Video Based Biometric Person Authentication*, pp. 964–974, Guildford, UK, June 2003.

[8] R. Paredes, J. C. Pérez, A. Juan, and E. Vidal. Local representations and a direct voting scheme for face recognition. In *Pattern Recognition in Information Systems*, pp. 71–79, Setúbal, Portugal, July 2001.

[9] M. Reinhold, D. Paulus, and H. Niemann. Appearance-based statistical object recognition by heterogeneous background and occlusions. In *Pattern Recognition, 23rd DAGM Symposium*, LNCS 2191, pp. 50–58, Munich, Germany, Sept. 2001.

[10] P. Simard, Y. Le Cun, and J. Denker. Efficient pattern recognition using a new transformation distance. In *Advances in Neural Information Processing Systems*, volume 5, pp. 50–58, San Mateo, CA, USA, 1993.