Adaptive Performance-Based Classifier Combination for Generic Object Recognition

Frank Mattern[1] *, Torsten Rohlfing[2], Joachim Denzler[1]

1) Lehrstuhl Digitale Bildverarbeitung
Friedrich-Schiller-Universität Jena
Ernst-Abbe-Platz 2, 07743 Jena, Germany
e-mail: {mattern,denzler}@informatik.uni-jena.de
2) Neuroscience Program, SRI International
333 Ravenswood Ave, Menlo Park,
CA 94025-3493, USA
e-mail : rohlfing@ieee.org

Abstract

It is well-established in the pattern recognition community that the performance of classifiers can be greatly improved by combining the outputs of multiple classifiers. In this paper, we introduce the concept of adaptive performance-based classifier combination, i.e., the weighting of classifiers based on their estimated recognition performance, to generic object recognition. Using an expectation-maximization (EM) algorithm previously applied to image segmentation, we learn the characteristics of individual generic object recognition classifiers. Using the ETH-80 data sets we demonstrate that by incorporating these performance estimates in a Bayesian classifier combination, the recognition rate of the combined classifications improves substantially over feature combination as well as simple and confidence-based voting. The EM algorithm has no tunable parameters and does not require a pre-classified training set during the learning stage. We conclude that adaptive performance-based classifier combination is a valuable and versatile tool to improve the performance of generic object recognition systems.

1 Introduction

Categorization of objects is an important ability of the human brain. Young children start very early to distinguish basic level categories [4] such as `dog`, `chair` or `ball`. Humans can also easily assign unknown objects to trained categories and can distinguish between known and unknown objects [1, 7]. According to psychological essentialism, objects with internal, possibly unknown, essences form a category [6], i.e., all existing chairs have something in common. Generic object recognition systems try to extract the features which represent the internal essences of the objects' categories. The generic object recognition task is complicated by the requirement to not only *recognize* known objects, but to *categorize* unknown objects, and to decide whether a particular object is already known or not. The discrimination is necessary, for example, to distinguish between tasks to search for a generic object of a certain category (e.g., "find *a* cup"), or to identify a specific object that was learned before (e.g., "find *my* cup").

It is quite difficult to find adequate features for generic object recognition because different features may be relevant for different generic classes. For example, color should not be relevant for the category `ball` but is probably helpful for the category `tomato` because certain colors are likely for tomatoes. Therefore, a combination of multiple classifiers, each targeting different features, promises to be more successful at generic object recognition than any single classifier could be.

Based on this observation, we start our work in this paper with a simple voting method and improve upon it by weighting the classifiers with a static confidence measures. As our core contribution, we then go on to show that further improvement can be achieved using a self-training adaptive approach for performance-based classifier combination. To this end, we apply a method named "Simultaneous Truth and Performance Level Evaluation" (STAPLE) that was introduced by Warfield

*This work was funded by the German Science Foundation (DFG) under grant DE 735/2–1.

et al. [16]. The algorithm weights classifiers based on their estimated recognition performances. It is adaptive, parameterless, and can be trained with or without ground truth data.

STAPLE has previously been applied to combination of image segmentations [15], an application in which there is no training set and no need for generalization beyond the current image. In this paper, we evaluate STAPLE's application for the generic object recognition, which is a substantially different problem. Using a standard database for generic object recognition, we classify objects with PCA-based feature combination, voting, a static confidence-weighted combination method, and STAPLE.

The remainder of this paper is organized as follows. Section 2.1 motivates the choice of classifier used in this paper. Section 2.2 categorizes the used voting and the STAPLE approach. In Section 2.3, we review the multi-class STAPLE algorithm and describe its application for performance-based classifier combination. The performed experiments and their results are described in Section 3. We conclude the paper in Section 4 with a discussion of our results and an outlook on future research.

2 Background and Methods

2.1 PCA-based generic object recognition

Unlike early approaches for generic object recognition that used geometric modeling (e.g., [14]), we extract features directly from the training images. These features are then used to learn a model for the specific object and the generic object category. First encouraging results for this approach to categorize objects based on appearance information extracted using principal component analysis (PCA) was presented by Jain *et al.* [9]. They built a hierarchy of categories and constructed a PCA-based model for each category. Image-to-model comparison was done using reconstruction error.

In [2] we studied the generic modeling and classification with mixtures of principal component analysis. Therein, generic object categorization was done by a unsupervised learning step using the EM algorithm, and the probabilistic model is used for Bayesian classification. In [3] we combined the approach with a hierarchical structure and examined the generic classification rate in a hierarchy. In [13] we examined segmentation-free generic classification methods with supervised training using nearest neighbor, mixtures of principal component analysis and kernel principal component methods. We found that on the generic data set nearest neighbor classification on PCA projected vectors worked best.

In [11] a generic image database was introduced, several appearance-based and contour-based approaches were compared, and a multi-cue combination was suggested. Leibe & Schiele [12] combined object categorization and segmentation and handled objects in real world scenes. LeCun *et al.* [10] created a large data set with 50 objects of 5 categories and 194,400 images. They tested nearest neighbor methods, support vector machines and convolution networks using raw pixel values and PCA-derived features. Most recent approaches also describe generic objects by local appearance-based methods [5] which recognize objects by their parts.

As PCA-based features on image gray values have worked well in the past, we apply in this paper PCA on different preprocessed images. We then examine if the method can further be improved. If the features are not mutually dependent (i.e., if their diversity is high), its is reasonable to expect that a combination of classifiers can further improve the recognition rate.

2.2 Classifier combination methods

The combination of classifiers to improve classification results is a popular research topic. Jain *et al.* [8] give a good overview of different classifier combination methods. They distinguish the combination methods based on whether they are trainable or adaptive, and also based on what type of information they use. The type of information that a classifiers can use consists of class labels, vectors of class ranks, or a confidence value for each class. The intuitive combination of classification results is the voting method. Each classifier can vote for one class, and the class with the most votes wins (see Eq. 5). The voting method combines only class labels, which is described in [8] as the "abstract level". Weighting each class which a confidence value and assigning as the result the class with the maximum combined confidence is a method that works on the "confidence level" (or "measurement level"). Both methods are not trainable and not adaptive, that is, they cannot be tuned to a specific classification problem, and they do not learn as they proceed. In the next section, we review the STAPLE [16] algorithm, an approach that is both trainable and adaptive.

2.3 Review of the multi-class STAPLE algorithm

The STAPLE algorithm was introduced by Warfield *et al.* [16] for ground truth estimation from multiple expert segmentations, and is based on classifiers which only provide abstract class labels. The algorithm was originally limited to a two class problem but was extended to the multi-class problem by Rohlfing *et al.* [15]. Below, we briefly review the multi-class STAPLE algorithm and introduce the notation as it was used in [15].

If a sample $\boldsymbol{x}$ is classified by the kth classifier in class j of L classes this will be described by

$$e_k(\boldsymbol{x}) = j \ , \tag{1}$$

where $j = 1 \dots L{+}1$. A rejected sample is assigned to pseudo class $L+1$. The fact (i.e., ground truth) that $\boldsymbol{x}$ is actually in class i is expressed by

$$\boldsymbol{x} \in C_i \ , \tag{2}$$

where $i = 1 \dots L$. Therefore, the conditional probability that the classifier k assigns the sample $\boldsymbol{x}$ to class j, whereas in fact it is in class i, can be described by

$$P(e_k(\boldsymbol{x}) = j | \boldsymbol{x} \in C_i) \ . \tag{3}$$

Using the above definitions, a general combination rule can be described. We seek as the output of the combined classifier $E(\boldsymbol{x})$ the class that maximizes the probability given a performance model M and all classifier decisions $e_k(\boldsymbol{x})$, where $k \in 1 \dots K$:

$$E(\boldsymbol{x}) = \arg\max_i P(x \in C_i | \forall_k e_k(\boldsymbol{x}), M) \ . \tag{4}$$

The simplest non-adaptive performance model is to assume equal recognition rates of all classifiers for all classes, which leads to the vote rule as mentioned in Section 2.2. A combined classifier $E_{\text{vote}}(\boldsymbol{x})$ can be easily constructed by

$$E_{\text{vote}}(\boldsymbol{x}) = \arg\max_i \sum\nolimits_k \begin{cases} 1 & \text{if } e_k(\boldsymbol{x}) = i, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

If the performance model of each classifier is known and the classifiers are conditionally independent of each other, then one can use the Bayes formula to calculate the combined classifier result by Bayesian inference as

$$E_{\text{bayes}}(\boldsymbol{x}) = \arg\max_i \frac{P(\boldsymbol{x} \in C_i) \prod_k P(e_k(\boldsymbol{x}) | \boldsymbol{x} \in C_i, M)}{\sum_j P(\boldsymbol{x} \in C_j) \prod_k P(e_k(\boldsymbol{x}) | \boldsymbol{x} \in C_j, M)}, \tag{6}$$

where $P(e_k(\boldsymbol{x}) | \boldsymbol{x} \in C_i, M)$ is the probability that classifier k in fact assigns the sample $\boldsymbol{x}$ to class $e_k(\boldsymbol{x})$. This is determined by the performance model M.

In the STAPLE algorithm, the performance model is based on Bayesian classifiers. The performance of each classifier k is described by its confusion matrix $N^{(k)}$, which is a $(L \times L{+}1)$-dimensional matrix. It contains as its elements $n_{i,j}^{(k)}$ the number of samples which belong in class i and are classified into class j, i.e.,

$$n_{i,j}^{(k)} = \#\{\boldsymbol{x} | \boldsymbol{x} \in C_i \wedge e_k(\boldsymbol{x}) = j\} \ . \tag{7}$$

The number of samples $\boldsymbol{x}$ in a class i is described by

$$n_{i,\cdot}^{(k)} = \#\{\boldsymbol{x} | \boldsymbol{x} \in C_i\} = \sum\nolimits_j n_{i,j}^{(k)} \ . \tag{8}$$

If the confusion matrix $N^{(k)}$ is given, then the conditional probability can be easily calculated by

$$P(e_k(\boldsymbol{x}) = j | \boldsymbol{x} \in C_i, N^{(k)}) = \frac{n_{i,j}^{(k)}}{n_{i,\cdot}^{(k)}} \ . \tag{9}$$

These conditional probabilities are the parameters of our performance model. For notational convenience we use $\lambda_{i,j}^{(k)} := n_{i,j}^{(k)} / n_{i,\cdot}^{(k)}$. For a given set of classifiers, the coefficients $\lambda_{i,j}^{(k)}$ are estimated using an EM algorithm as follows.

In the expectation step, weights $W_i(\boldsymbol{x})$ are calculated, which correspond to the probability that the sample $\boldsymbol{x}$ belongs to class i. Assuming the conditions for Eq. 6 hold and no *a priori* knowledge is available, we get:

$$\begin{aligned} W_i(\boldsymbol{x}) &= P(x \in C_i | e_1(\boldsymbol{x}), \dots, e_K(\boldsymbol{x}), M) \\ &= \frac{\prod_k \lambda_{i,e_k(\boldsymbol{x})}^{(k)}}{\sum_j \prod_k \lambda_{i,e_k(\boldsymbol{x})}^{(k)}}. \end{aligned} \tag{10}$$

The maximization step is performed by calculating

$$\hat{\lambda}_{i,j}^{(k)} = \frac{\sum_{\boldsymbol{x}: e_k(\boldsymbol{x}) = j} W_i(\boldsymbol{x})}{\sum_x W_i(\boldsymbol{x})} \ . \tag{11}$$

We initialize the EM algorithm with the voting method as follows:

$$n_{i,j}^{(k)} = \#\{\boldsymbol{x} | E_{\text{vote}}(\boldsymbol{x}) = i \wedge e_k(\boldsymbol{x}) = j\} \ . \tag{12}$$

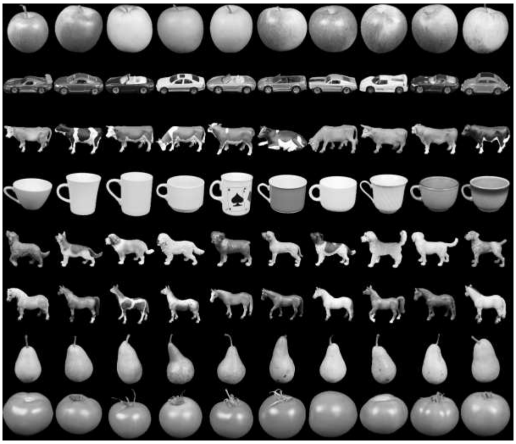


Figure 1: Sample images of the objects of the ETH-80 database [11]

3 Experiments

In this section, different features are examined for the use in generic object recognition. We investigate the improvement achieved by the combination of the features as well as combinations of single-feature classifiers. We describe below the benchmarking environment used for evaluation and present the results of the different methods.

3.1 Benchmarking Environment

To evaluate all classification and combination methods, we use the ETH-80 database [11], which contains 80 objects from 8 categories: `apple`, `car`, `cow`, `cup`, `dog`, `horse`, `pear` and `tomato` (see Figure 1). Each object is represented by 41 images of views from the upper hemisphere. The experiments are performed using 128×128 pixel images, with each image cropped close to the object boundaries. To generate comparable results to Leibe & Schiele [11], the test is performed by cross-validation with a leave-one-object-out strategy. One of the 80 objects is not trained. This "unknown" object must accordingly be classified into the correct object category.

The images of the database are processed with different feature transformation approaches. We used methods that emphasize color, shape, edge, or frequency information. Figure 2 illustrates some of the features. In (2a) the RGB color channels and intensity gray value image are shown, which are used in the "color" feature, and (2b) shows the hue component of the HSV model which is declared as the "hue" feature. In (2c) the images are generated by the inverse Euclidean distance transformation of the objects shape. Each pixel grayvalue is assigned to the value of the distance to the nearest non-feature pixel. We used the Fourier transformed features "fft" in (2d) using the absolute values of the parameters and also Haar wavelet transformed features. For feature generation, we used the simple scaling function

$$\phi(t) = \begin{cases} 1 & : \quad 0 < t < 1 \\ 0 & : \quad \text{otherwise} \end{cases} \tag{13}$$

with the associated Haar wavelet:

$$\psi(t) = \begin{cases} 1 & : \quad 0 < t < 0.5 \\ -1 & : \quad 0.5 < t < 1 \\ 0 & : \quad \text{otherwise} \end{cases}$$

$$h_0 = \frac{1}{\sqrt{2}}, \quad h_1 = -\frac{1}{\sqrt{2}} \tag{14}$$

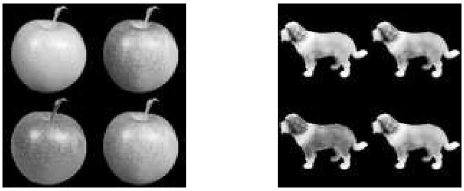
From the decomposition vector of hierarchy level one we use only a selection of low-pass and first horizontal high-pass coefficients.

All resulting image vectors of the training set are used to calculate a PCA transformation. The 100 eigenvectors which correspond to the largest eigenvalues are used to form an eigenspace projection matrix. Previous experiments [13] corroborate that 100 dimensions are a good choice between space and time consumption and recognition rate. As a model, the projected training vectors in that eigenspace are used. The nearest neighbor (NN) classification is used to assign the category to a test image. For evaluation, the global recognition rate for each object category is calculated.

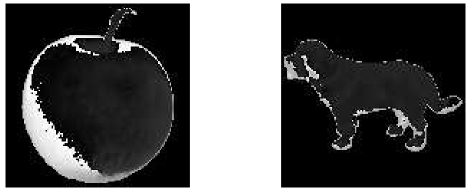
As mentioned before, a combination of different classification methods can often improve the generic recognition rate. There are two possible approaches to combine the features. The so-called *early combination* collects all features and selects the most significant features to construct a classifier. In a *late combination* approach, on the other hand, each feature set is classified by a separate classifier and the results are combined. We examined the first approach using PCA for feature selection of the combined features. The second approach was realized with the STAPLE [15] algorithm and with the voting method for combining the classifiers.

3.2 Results with Single Feature Selection

For comparison, the NN classifier is applied to original images and to the images which are resam-

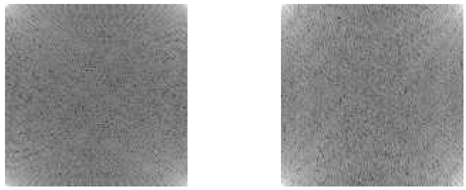


(a) Color RGB channels and gray values



(b) Hue values of the image

(c) Inverted images of Euclidean distance transform



(d) Fourier transformed images using absolute values

Figure 2: (a) Samples of training images apple (left) and dog (right). (b)–(d) Different features extracted from the images in (a).

		scaled images				
	PCA	8x8	16x16	32x32	64x64	128x128
apple	78.0%	80.2%	79.5%	79.5%	79.3%	77.1%
car	99.5%	89.3%	99.8%	99.5%	99.5%	98.8%
cow	69.8%	55.6%	67.1%	67.6%	62.7%	60.0%
cup	98.8%	92.7%	99.0%	98.3%	97.1%	95.4%
dog	58.8%	42.4%	58.3%	57.3%	57.3%	54.9%
horse	73.7%	56.3%	75.6%	75.9%	76.3%	77.1%
pear	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
tomato	90.0%	94.6%	89.5%	90.2%	92.2%	93.4%
overall	83.6%	76.4%	83.6%	83.5%	83.0%	82.1%

Table 1: Result of generic object recognition by NN classification using PCA dimensionality reduction to 100 compared to NN classification directly on the gray values of scaled images.

pled to 8×8, 16×16, 32×32, and 64×64 pixels with bicubic interpolation. In Table 1 we see how the generic object recognition works with NN classifiers using the original image and their resampled versions. Using the full gray-value image for NN classifiers, the ETH-80 dataset cannot be categorized without errors. The scaled images improve the original classification rate. The categories `dog`, `cow` and `horse` are not well represented by the model, whereas `pear`, `car`, `cup` and also `tomato` can be quite well separated. Most of the apples are correctly and consistently classified into the generic class `apple`. Object apple10 is always classified as `tomato` and drags down the classification rate. The best overall result, a recognition rate of 83.6% across all classes, is obtained using 16×16 pixel images.

Next, instead of scaling the images themselves, we applied a PCA transformation and used the 100-dimensional projections of the training images into the eigenspace for NN classification as mentioned in Section 3.1. The generic recognition rate is 83.6% and, therefore, equally good as the best achieved by resampling (see Table 1), but using a substantially smaller number of features. We note that the recognition rates of cars, cups and pears are consistently near 100%, while cows and dogs have very low recognition rates.

We experimented with several feature selection methods. For demonstrating the effect we used a set of feature selection methods which works best on the dataset. We noticed that color information is helpful for this database. Using color images instead of gray level images the generic recognition rate increases to 85.2% (see Table 2). This is not obvious because color information can also cause error by limiting the ability of the classifier to generalize, e.g., when classifying a yellow tomato. Taking only the hue component of the HSV model provides a worse overall generic recognition rate, but the best recognition rate for apples. Especially for generic object recognition, the shape is in general a reliable feature. Using only the distance transformation based on the object shape, the object recognition rate is slightly improved to 85.6%. But this assumes that we have an effective shape detection algorithm. Frequency-based features can also improve the result. But the Fourier transformation (FFT) using the absolute values fails for generic object recognition. However, wavelet features improve the recognition rate up to 87.3%, because the misclassification rate of dogs and cows are considerably reduced (see Table 2).

The results are comparable to the methods of Leibe & Schiele [11]. The PCA approach with one eigenspace and without preprocessing achieved a recognition rate of 83.6%, whereas Leibe with a multi-eigenspace approach achieved only 83.0% using the same data sets. Leibe & Schiele [11] achieved the best results (recognition rate 86.4%) with contour-based approaches. This can be further improved with the wavelet-based approach.

3.3 Results with Classifier Combination

Combination of the features and feature selection with PCA did not work well. We combined the gray values with different features, including edge images, distance transformation, FFT, and wavelet transformation. We then used a PCA transformation to select important features. Feature combination did not improve the recognition rate. The best result (84.8%) was obtained by combining wavelet features and gray values (see Table 2).

The voting approach improved the recognition rate to 88.2%. The two categories `car` and `cup` were perfectly recognized in the combination, even though none of the individual classifiers alone achieved perfect recognition result on the cups. Using confidence values for the classification, this could be improved. A confidence value is defined as $C(\boldsymbol{x}) \in [0:1]$ where $C(\boldsymbol{x}) = 1$ expresses that the classifier is 100% sure that the classification is correct and $C(\boldsymbol{x}) = 0$ means that the classifier is entirely uncertain. In the case of nearest neighbor classification an adequate reasoning might be as follows. If the distance to the to the nearest neighbor d_{nn} is zero, then the confidence measure should be one. If, on the other hand, $d_{\text{nn}} = d_{\text{sn}}$, where d_{sn} is the distance to the second best class, then the classifier cannot make a deterministic decision because two classes are essentially equally likely. We experimented with several functions that satisfy these conditions The best results, were achieved with the confidence measure

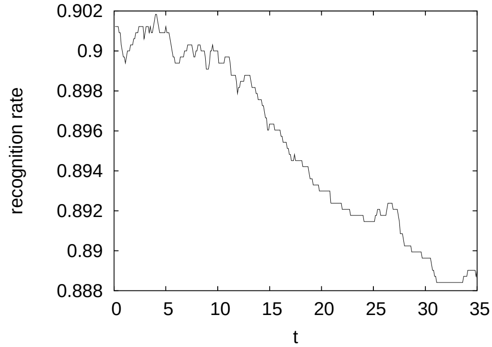


Figure 3: Recognition rate of the combined confidence-based classification vs. parameter t of the confidence weighting function $C_t(\boldsymbol{x})$.

$$C_t(\boldsymbol{x}) = 1 - \exp\left(t\frac{d_{\text{nn}}(\boldsymbol{x}) - d_{\text{sn}}(\boldsymbol{x})}{d_{\text{nn}}(\boldsymbol{x})}\right). \quad (15)$$

This function has a free parameter, t. Using $t = 4.0$ we achieved a recognition rate of 90.2%. Note that using this confidence measure, one has to choose the right parameter to achieve good results. Using a "bad" parameter, performance suffers substantially, as is illustrated in Figure 3.

Classifier combination using the STAPLE algorithm outperformed all other individual classifiers and the feature combination method with a generic recognition rate of 90.6%. The differences are statistically significant with $p < 10^{-4}$. The best achievable recognition rate using the correct confusion matrix instead of the estimated one is 91.7%. Even though we applied the STAPLE method to abstract labels in the present work, it performed slightly better than the constant weighting approach that operates on the confidence level.

The results of this approach are comparable to the multi-cue combination approach without contours of Leibe & Schiele [11] which achieves 90.0%.

	features				
	color	edt	fft	hue	wt
apple	80.0%	72.9%	80.5%	**87.3%**	82.0%
car	99.5%	**100.0%**	93.2%	95.6%	**100.0%**
cow	71.2%	**86.8%**	30.2%	60.5%	82.2%
cup	99.5%	**99.8%**	90.7%	90.5%	**99.8%**
dog	59.5%	75.4%	68.0%	50.5%	**78.0%**
horse	75.6%	**80.7%**	25.6%	43.2%	77.1%
pear	**100.0%**	**100.0%**	79.0%	99.8%	99.8%
tomato	**96.6%**	69.0%	38.0%	88.8%	80.0%
overall	85.2%	85.6%	63.2%	77.0%	**87.3%**

(a)

	feature combination	classifier combination		
	wtcom	vote	vote with co	staple
apple	75.4%	**85.4%**	83.4%	83.4%
car	99.8%	**100.0%**	**100.0%**	**100.0%**
cow	70.7%	84.1%	82.0%	**84.6%**
cup	99.0%	**100.0%**	**100.0%**	**100.0%**
dog	66.6%	**79.5%**	77.1%	78.8%
horse	78.3%	72.0%	82.4%	**84.4%**
pear	**100.0%**	99.8%	**100.0%**	**100.0%**
tomato	89.0%	84.9%	88.0%	**93.9%**
overall	84.8%	88.2%	90.2%	**90.6%**

(b)

Table 2: Recognition rates using NN classification of PCA-transformed feature vectors and combinations of features and classifications. The best recognition rate in each row is printed in bold face. (a) Feature vectors were extracted from the original color images (color), the hue of the HSV model (hue), generated by FFT (fft), Haar wavelet transformation (wt) and by Euclidean distance transformation based on image shape (edt). (b) Feature combination of gray value images and wavelet features (wtcom) and combined generic recognition results using voting, voting with confidence measure (co) and STAPLE.

4 Conclusion and Outlook

We have shown that feature selection methods can improve the generic classification rate. Color information improves the recognition rate for the ETH-80 dataset which may be caused by the choice of objects. FFT based features fail whereas wavelet based features performs best. Early combination of features and selection with PCA does not improve the recognition rate.

The overall system recognition rate can be improved over the performances of the individual classifiers by using voting and especially voting approaches that incorporate a classifier performance model, such as the STAPLE algorithm. We achieved 90.6% recognition rate by combining 5 classifiers using STAPLE, where each had an individual recognition rate between 63.2% and 87.0%. Which features are vital for the recognition of a category depends on the categories. Color, for example, can provide essential evidence for a particular class, while for a generic category it is often irrelevant. Frequently, object shape is an important indicator for a category, but shape alone is not always sufficiently specific to a single class.

The multi-class STAPLE algorithm outperformed all individual feature classifiers, as well as all classifier and feature combination methods. While its superiority over voting with a confidence function is not statistically significant, we point out that STAPLE is substantially more universal and potentially robust. While we had to tune the parameter t of the confidence weighting function C_t to achieve optimum performance, no such tuning was necessary for STAPLE. Consequently, we have less reason to expect STAPLE's performance to suffer from bad parameter choices (let alone the choice of a bad confidence function).

In terms of versatility, STAPLE can be trained on pre-classified samples, which improves the recognition rate. It can also be re-trained during the classification phase without requiring pre-classified data, which makes it capable of classifying objects adaptively. Unlike all other methods evaluated here, it requires no training set since it estimates the ground truth in the process of adjusting the classifier performances. STAPLE only requires, and indeed performs very well, on abstract class label information, but its performance may further improve when using confidence values as its input. This will be evaluated in future research.

References

[1] D. Bovet, J. Vauclair, and A. Blaye. Categorization and abstraction abilities in 3-year-old children: a comparison with monkey data. *Animal Cognition*, 8(1):53–59, 2005.

[2] C. Drexler, F. Mattern, and J. Denzler. Appearance Based Generic Object Modeling and Recognition Using Probabilistic Principal Component Analysis. In *Proc. of the 24th DAGM Symposium for Pattern Recognition*, pp. 100–108, Zurich, 2002.

[3] C. Drexler, F. Mattern, and J. Denzler. Generic Hierarchic Object Models and Classification based on Probabilistic PCA. In *IAPR Workshop on Machine Vision Applications*, pp. 435–438, Nara, Japan, 2002.

[4] W.D. Gray, D.M. Johnson, E. Rosch, C.B. Mervis, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, 1976.

[5] L. Fei-Fei, R. Fergus, and P. Perona. Learning Generative Visual Models From Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. In *IEEE CVPR Workshop on Generative Model Based Vision (WGMBV)*, 2004.

[6] S.A. Gelman. Psychological essentialism in children. *Trends in Cognitive Sciences*, 8(9):404–409, 2004.

[7] S.A. Gelman and E.M. Markman. Young children's inductions from natural kinds: the role of categories and appearances. *Child Development*, 58(6):1532–1541, 1987.

[8] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[9] G. Jain, M. Agarwal, R. Choudhury, and S. Chaudhury. Appearance-based Generic Object Recognition. In *Proc. Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 20–22, Bangalore, India, 2000.

[10] Y. LeCun, F.J. Huang, and L. Bottou. Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting. In *Proc. of International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 97–104, Washington, D.C., 2004.

[11] B. Leibe and B. Schiele. Analyzing Appearance and Contour Based Methods for Object Categorization. In *Proc. of International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 409–415, Madison, WI, 2003.

[12] B. Leibe and B. Schiele. Combined Object Categorization and Segmentation with an Implicit Shape Model. In *ECCV Workshop on Statistical Learning in Computer Vision*, Prague, 2004.

[13] F. Mattern and J. Denzler. Comparison of Appearance-Based Methods for Generic Object Recognition. *Pattern Recognition and Image Analysis*, 14(2):255–261, 2004.

[14] G. Medioni and A. Francois. 3-D Structures for Generic Object Recognition. *Proc. of the IAPR International Conference on Pattern Recognition*, 1:30–37, 2000.

[15] T. Rohlfing, D.B. Russakoff, and C.R. Maurer, Jr. Performance-Based Classifier Combination in Atlas-Based Image Segmentation Using Expectation-Maximization Parameter Estimation. *IEEE Transactions on Medical Imaging*, 23(8):983–994, 2004.

[16] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[17] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3), pp. 418–435, 1992.