A Statistical Framework for Multi-Object Recognition

Daniel Keysers, Jörg Dahmen, Hermann Ney, and Mark Oliver Güld
keysers@cs.rwth-aachen.de

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology, D-52056 Aachen, Germany

Supervisor: Prof. Dr.-Ing H. Ney
Type: PhD thesis
GI subjects: image understanding (1.0.4), machine learning (1.1.3)

Abstract
In this paper, we present a statistical framework for the recognition of multiple objects in an image, which is a generalization of the Bayesian decision rule. The approach takes into account the interdependence of several operations including recognition and transformation of the objects and segmentation. We present first experimental results on single objects and on artificially generated scenes. Although the computational complexity of the approach is high, the additional effort seems justified and there is potential for reduction of complexity.

1 Introduction

In the past years, statistical methods have been successfully applied to the task of object recognition in images. In most cases, the methods used are either specialized on the recognition of single objects in a known position or on determining the position of a known object in the image. In this paper, we present a statistical framework that includes the approaches mentioned above as special cases and is designed for the *recognition of multiple objects* in an image. This task is strongly related to the recognition of objects in the presence of varying background. Usually, this type of recognition is performed in combination with a separate segmentation step, which is inherently error-prone. To carry out scene recognition, an automatic system is faced with the interdependence of several operations, including segmentation, object detection and recognition and transformations of the objects. Therefore, the main concept of the approach presented here is to generalize the classical Bayes' decision rule to more complex object recognition tasks, i.e. to choose among all possible object and background configurations the one that maximizes the probability that the image was produced by this configuration. Thus, a meaningful segmentation of the image is implicitly determined. The need for such a *holistic* approach to image recognition (i.e. the necessity to explain the whole image, integrating segmentation and recognition) has been recognized before [1]. It is also known that weaknesses of many current recognition systems are the reliance on a separate segmentation step and the removal of context information. These problems are illustrated in Fig. 1 and further problems include recognition of partially occluded and very close objects in an image. In spite of the increased computational complexity of the holistic approach, the success of the holistic paradigm in speech recognition, where interdependence between time alignment, word boundaries and syntactic constraints exist, shows that the additional effort may well be justified [2]. From the domain of automatic speech recognition, there are also various methods known to reduce the necessary amount of computations.

2 Classification of Single and Multiple Objects

In this section, starting from the classical Bayes' rule, we introduce the framework for multiple object recognition. In all the cases considered, it is important to model the variability of the image objects.

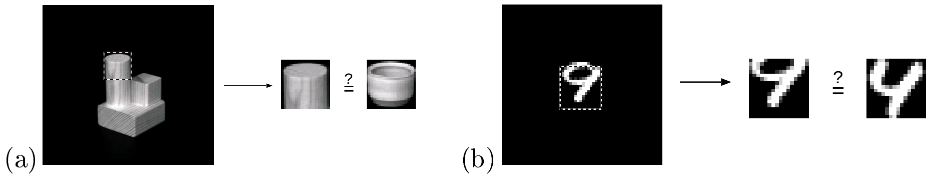

Figure 1: Problems with interdependence in recognition: (a) Only a small part of the original object is explained, possibly resulting in a misclassification (COIL-20 data). (b) Effect of small localization errors on the classification result (USPS data).

Although not presented in detail here, the methods used in the experiments to deal with variability are invariant distance measures and virtual training and test data [3, 4, 5, 6].
To classify an observation $x \in \mathbb{R}^D$ we use the Bayesian decision rule

$$x \longmapsto r(x) = \underset{k}{\operatorname{argmax}} \{p(k)p(x|k)\}, \tag{1}$$

where $p(k)$ is the prior probability of class k and $p(x|k)$ is the class-conditional probability for the observation x given class k [3]. For multiple object recognition, we extend the elementary decision rule into the following directions:

- We assume that the scene x contains an unknown number M of objects belonging to the classes $k_1, ..., k_M =: k_1^M$. Reference models $p(x|\mu_k)$ exist for each of the classes $k = 1, \ldots, K$, μ_0 representing background.
- We take decisions about object boundaries, i.e. the original scene is implicitly partitioned into $M + 1$ regions I_0^M, where I_m contains the m-th object and I_0 represents the background.
- The reference models may be subject to transformations (rotation, scale, translation, etc.). That is, given transformation parameters ϑ_1^M, the m-th reference is mapped to $\mu_{k_m} \rightarrow \tilde{\mu}(\mu_{k_m}, \vartheta_m)$.

The idea is now to consider all unknown parameters, i.e. M, k_1^M, ϑ_1^M and I_0^M and to search the hypothesis which best explains the given scene. Note that this means that any pixel in the scene must be assigned either to an object or to the background class. The resulting decision rule is:

$$r(\{x_{ij}\}) = \underset{M, k_1^M, \vartheta_1^M, I_0^M}{\operatorname{argmax}} \left\{ p(k_1^M) \prod_{(i,j) \in I_0} p_0(x_{ij}|\mu_0) \prod_{m=1}^{M} p_{k_m}(x_{I_m}|\tilde{\mu}(\mu_{k_m}, \vartheta_m)) \right\} \tag{2}$$

where $\{x_{ij}\}$ denotes the scene to be classified and x_{I_m} is the feature vector extracted from region I_m. To model the references $p_{k_m}(x_{I_m}|\tilde{\mu}(\mu_{k_m}, \vartheta_m))$, Gaussian kernel or mixture densities can be used, where invariance aspects can be directly incorporated into the model using probabilistic modelling of variability [3]. In (2), $p(k_1^M)$ is shorthand for a prior over the combination of objects in the scene, which may depend on the transformation parameters (e.g. a head above a body is more likely than vice versa). The consideration of all the components of the presented decision rule is a long-term goal. In the experiments performed so far, we have started with the consideration of the interdependence between segmentation and recognition.
Discussion of related work: Once the maximizing arguments in (2) have been determined, it is straightforward to construct a parse tree as a description of the image from the implicit segmentation information, which is done using a neural net based approach in [1]. Special attention to the subject of occlusion is payed in [7], but in this work mainly object contours are considered for recognition, not the objects themselves. Some considerations with respect to a statistical model for multiple images can also be found in [8]. Here, the author concentrates on determining the unknown (3D) transformation parameters in the recognition process as well as improving feature extraction. It is shown that localization can be improved by explicit modelling of the background, although no global optimization is performed in the experiments. Note that the presented framework imposes no restriction on the type of feature extraction used. In the experiments presented in the following, we make use of *appearance based* pattern recognition, i.e. each pixel of an image is interpreted as a feature.

3 Single Object Recognition Experiments

First, we present results for the statistical approach restricted to $M = 1$, i.e. it is known that each image contains exactly one object. If we impose the further restriction $I_0 = \emptyset$ (no background present), we arrive at the classical Bayesian decision rule including an explicit model of variability. As previous experiments have shown, this approach achieves state-of-the-art results in recognition tasks as optical character recognition (experiments on the well known USPS and MNIST databases yielding error rates of 2.2% and 1.0% respectively [3, 4]).
Tasks in which this restriction is not appropriate, i.e. tasks which require a possibly non-empty background region can also be characterized as recognition of a single object including localization. To illustrate this case, we present results obtained on two image databases, starting with the collection of radiographs from the IRMA project (Image Retrieval in Medical Applications [9]). The database consists of medical radiograph images taken from daily routine, which are secondary digital. The sizes

Figure 2: Example radiographs taken from the IRMA database, scaled to common height, left to right: abdomen, limbs, breast, skull, chest and spine.

of the anonymized images range from about 200×200 pixels to about 2,500×2,500 pixels. The corpus consists of 110 abdomen, 706 limbs, 103 breast, 110 skull, 410 chest and 178 spine radiographs, summing up to a total of 1,617 images (cp. Fig. 2). To speed up the classification process, the original images are scaled down to a common height of 32 pixels (keeping the original aspect ratio). Since there are only 1,617 images available, we make use of a leaving-one-out approach. That is, to classify an image we use the remaining 1,616 images as references in a Gaussian kernel density $p_{k_m}(x_{I_m}|\tilde{\mu}(\mu_{k_m}, \vartheta_m))$ with class specific diagonal covariance matrices. The prior probabilities $p(k)$ are modeled using relative frequencies. The background model used so far is simple, assuming a constant background grayvalue. Furthermore, a penalty term is introduced, which is based on the different sizes of observation and reference image (preferring images of roughly the same size). The obtained IRMA results are summarized in Table 1 and compared to other available results.

The second database used in the experiments is the well-known Columbia University Object Image Library (COIL-20), which consists of images taken from 20 different 3D-objects viewed from varying positions. Each image contains a single object (which is subject to different illumination conditions). There are 1,440 reference images of size 128×128 pixels available (called *processed* data), as well as 360 test images of size 448×416 pixels (called *unprocessed* data). To guarantee that training and test set are sufficiently different, only images with odd rotation angle are used as references and only images with even rotation angle as test scenes. Thus, a number of 720 reference images and 180 test images remains. It should be noted that by doing so, an observation to be classified always differs from the optimal reference by five degrees. This is contrary to the experiments conducted in [5], where the test scenes (which are unavailable) differed by 2.5 degrees in the worst case. Thus, the experiments conducted throughout this work can be regarded to be a harder classification task. To speed up the recognition process, the reference images were scaled down to 24×24 pixels. On the COIL-20 data, an error rate of 0% was thus obtained using 40 logarithmic scale levels for the scale transformation and pruning techniques based on the given, black background in the images. It should be noted that other research groups split the COIL-20 processed data into a training and a test set. Using this splitting, the problem can be treated as a single object recognition problem without localization (as training and test images are of the same size) and even a simple 1-nearest neighbor classifier yields an error rate of 0%.

4 Multiple Object Recognition Experiments

In this section, we present first results for the recognition of multiple objects in an image using the presented statistical approach. The main difficulty here is that there is no widely used database available that is suited for this task, which makes it difficult to compare obtained results. Therefore, we constructed artificial data by randomly positioning multiple digits of the first 100 test images of the USPS corpus in varying size (16×16 to 32×32) and position inside an image of size 96×96 using black background. Note that e.g. in [1] artificial data based on handwritten digits was used as well. Three examples of test scenes and recognized references are shown in Fig. 3(a). The resulting recognition obtained an error rate of 6%, which is the same as for a 1-NN classifier on the original 100 test images despite the additional uncertainty about position and scale of the digits. The algorithm used kernel densities for the class specific densities, a uniform distribution for $p(k_1^M)$ and 10 logarithmic scale levels for the scale transformation. Furthermore, knowledge about the constant background was

Table 1: Summary of results for the IRMA database (error rate [%]).

method	ER [%]
kernel densities, Mahalanobis distance	14.0
+ invariance (tangent distance, local threshold, image distortion [3])	8.2
square images, 1-nearest-neighbor	18.2
cooccurrence matrices	29.0

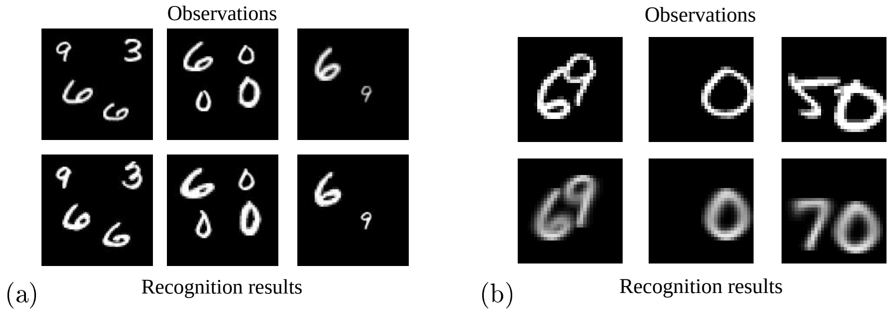

Figure 3: Examples of multi-object recognition. (a) Using knowledge about minimum distance and (b) straightforward implementation of (2) with the ability to recognize very close objects.

used, i.e. the background model is assumed to be a Gaussian distribution with zero mean and unit variance. Note that by simply relying on local decisions, the error rate significantly deteriorates to over 40% (even to 73% for a 1-NN) due to problems as the one illustrated in Figure 1(b).
To overcome the problems with very close objects, a key experiment was conducted using a straightforward implementation of (2) using hypothesis generation. Here, the original US Postal Service digits were randomly placed in a 32×32 pixels sized scene (with no scale variations). In order to reduce the computational complexity, a single density model was used for the references, while background and prior models were not changed. Examples of the resulting recognitions are shown in Figure 3(b). Note that now adjacent objects can be handled, too.

5 Conclusion

In this paper, we presented a statistical framework for the multi-object recognition, which is a valid generalization of the classical Bayes' decision rule. This holistic approach allows to integrate segmentation and classification so that the segmentation of an image can be obtained in a statistically reasonable way. First results on standard databases for single objects and on artificially created data for multiple objects seem very promising. The computational complexity of the approach is high in comparison to other heuristic methods, but is justified by the increased recognition performance. Furthermore, there exist techniques to reduce the computational cost (like e.g. beam search) that have not yet been fully incorporated. Current work includes improving the background model, which should be explicitly learned, and extension of the approach to the training phase, i.e. learning of the object model from a scene.

References

[1] G. Hinton, Z. Ghahramani, and Y. Teh. Learning to Parse Images. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*. MIT Press, pages 463–469, 2000.

[2] H. Ney and S. Ortmanns. Progress in Dynamic Programming Search for LVCSR. *Proceedings of the IEEE*, 88(8):1224–1240, August 2000.

[3] J. Dahmen, D. Keysers, H. Ney, and M. O. Güld. Statistical Image Object Recognition using Mixture Densities. *Journal of Mathematical Imaging and Vision*, 14(3):285–296, May 2001.

[4] D. Keysers, J. Dahmen, T. Theiner, and H. Ney. Experiments with an Extended Tangent Distance. In *Proceedings 15th International Conference on Pattern Recognition*, volume 2, Barcelona, Spain, pages 38–42, September 2000.

[5] H. Murase and S. Nayar. Visual Learning and Recognition of 3-D Objects from Appearance. *International Journal of Computer Vision*, 14(1):5–24, January 1995.

[6] P. Simard, Y. Le Cun, and J. Denker. Efficient Pattern Recognition Using a New Transformation Distance. In S. Hanson, J. Cowan, and C. Giles, editors, *Advances in Neural Information Processing Systems 5*. Morgan Kaufmann, pages 50–58, 1993.

[7] K. Mardia, W. Qian, D. Shah, and K. de Souza. Deformable Template Recognition of Multiple Occluded Objects. *IEEE Trans. Pattern Analysis Machine Intelligence*, 19(9):1035–1042, September 1997.

[8] J. Pösl. *Erscheinungsbasierte statistische Objekterkennung*. PhD thesis, Universität Erlangen Nürnberg, Erlangen, 1998. Shaker Verlag, Aachen.

[9] T. Lehmann, B. Wein, J. Dahmen, J. Bredno, F. Vogelsang, and M. Kohnen. Content-Based Image Retrieval in Medical Applications: A Novel Multi-Step Approach. In *Proceedings SPIE*, volume 3972(32), pages 312–320, February 2000.