Training and Recognition of Complex Scenes using a Holistic Statistical Model

Daniel Keysers, Michael Motter, Thomas Deselaers, and Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology, D-52056 Aachen, Germany
{keysers, motter, deselaers, ney}@informatik.rwth-aachen.de

Abstract. We present a holistic statistical model for the automatic analysis of complex scenes. Here, holistic refers to an integrated approach that does not take local decisions about segmentation or object transformations. Starting from Bayes' decision rule, we develop an appearance-based approach explaining all pixels in the given scene using an explicit background model. This allows the training of object references from unsegmented data and recognition of complex scenes. We present empirical results on different databases obtaining state-of-the-art results on two databases where a comparison to other methods is possible. To obtain quantifiable results for object-based recognition, we introduce a new database with subsets of different difficulties.

1 Introduction

The increasing availability of digital images causes a growing interest in automatic classification of such images. Up to now, approaches to classification, indexing, or retrieval are usually not based on the objects present in the image, but mostly on features derived from color or texture. This is due to the fact that automatic segmentation of objects in presence of inhomogeneous background is still an unsolved problem [7]. Approaches to image object recognition rely on manually pre-segmented data for training. These algorithms also perform best for homogeneous or static background but ignoring background information in automatic recognition can cause classification errors.

In this paper we address the problem of automatically determining object references and object-based classification in the presence of background. We present an appearance-based holistic statistical model for automatic training and recognition of image objects that explicitly takes into account the image background. Starting from Bayes' decision rule, which is the best we can do to minimize the error rate, we avoid explicit segmentation and determination of transformation parameters but instead consider these as integral parts of the decision problem. This is done to avoid incorrect local decisions. This holistic approach takes into consideration experiences in speech recognition, where explicit segmentation of 'objects' (words) and background is neither done in training, nor in recognition. Note that treatment of distortions and transformations is computationally significantly more demanding in 2D (e.g. images) than in 1D (e.g. speech signals).

Related work. The problems addressed here have been considered by other authors with different approaches. We discuss two works that are closely related:

A statistical model for object recognition in the presence of heterogeneous background and occlusions was presented in [6]. The authors use wavelet features to determine the local probabilities of a position in the image belonging to an object or to the background. The background is modeled by a uniform distribution. The assumption of statistical independence of the object features is reported to produce best results. The problem of automatic training in presence of heterogeneous background is not addressed. The authors report 0% error rate on a classification and localization task, in the presence of rotation and translation.

A similar model to the one presented here has been independently proposed in [1]. The authors introduce transformed mixtures of Gaussians that are used to learn representations on different databases of image data. They provide a detailed description of the statistical model. They consider only translations for an image database with background but do not present quantifiable results for this case. Instead, they only compare the results to a Gaussian mixture not regarding transformations. Error rates are only given for a set of synthetic 9×9 images in comparison to Gaussian mixtures.

2 Statistical Model and Decision Making

Principles. To classify an observation $\mathbf{X} \in \mathbb{R}^{I \times J}$ we use Bayes' decision rule

$$\mathbf{X} \longmapsto r(\mathbf{X}) = \underset{k}{\operatorname{argmax}} \left\{ p(k)\, p(\mathbf{X}|k) \right\}, \tag{1}$$

where $p(k)$ is the prior probability of class k and $p(\mathbf{X}|k)$ is the class-conditional probability for the observation $\mathbf{X}$ given class k. For holistic recognition, we extend the elementary decision rule (1) into the following directions:

- We assume that the scene $\mathbf{X}$ contains an unknown number M of objects belonging to the classes $k_1, ..., k_M =: k_1^M$. Reference models $p(\mathbf{X}|\boldsymbol{\mu}_k)$ exist for each of the classes $k = 1, \ldots, K$, and μ_0 represents the background.
- We take decisions about object boundaries, i.e. the original scene is implicitly partitioned into $M + 1$ regions I_0^M, where $I_m \subset \{(i, j) : i = 1, ..., I, j = 1, ..., J\}$ is assumed to contain the m-th object and I_0 the background.
- The reference models may be subject to certain transformations (rotation, scale, translation, etc.). That is, given transformation parameters $\boldsymbol{\vartheta}_1^M$, the m-th reference is mapped to $\boldsymbol{\mu}_{k_m} \rightarrow \boldsymbol{\mu}_{k_m}(\boldsymbol{\vartheta}_m)$.

The unknown parameters $M, k_1^M, \boldsymbol{\vartheta}_1^M$ and (implicitly) I_0^M must be considered and the hypothesis which best explains the given scene is searched. This must be done considering the interdependence between the image partitioning, transformation parameters and hypothesized objects, where in the holistic concept partitioning is a part of the classification process. Note that this means that any pixel in the scene must be assigned either to an object or to the background

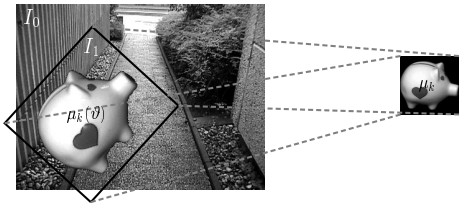


Fig. 1. Implicit partitioning and comparison during the search.

class. This model has been introduced in [3], where a restricted version was used in the experiments, only allowing horizontal shift. The resulting decision rule is:

$$r(\mathbf{X}) = \underset{M,k_1^M}{\operatorname{argmax}} \Big\{ \max_{\boldsymbol{\vartheta}_1^M} \big\{ p(\boldsymbol{\vartheta}_1^M) \cdot p(k_1^M) \cdot \prod_{m=0}^{M} p(\mathbf{X}_{I_m} | \boldsymbol{\mu}_{k_m}(\boldsymbol{\vartheta}_m)) \big\} \Big\}, \tag{2}$$

where $\mathbf{X}$ denotes the scene to be classified and $\mathbf{X}_{I_m}$ is the feature vector extracted from region I_m. Instead of performing a summation over the parameters $\boldsymbol{\vartheta}_1^M$, we apply the common maximum approximation here. Invariance aspects can be directly incorporated into the models chosen for the density functions using a probabilistic model of variability. In (2), $p(k_1^M)$ is a prior over the combination of objects in the scene, which may depend on the transformation parameters and the combination of objects.

Constraints. Regarding the components of the presented decision rule (2), we start with the consideration of the interdependence between segmentation and recognition. For the identification of one object in the presence of inhomogeneous background we assume $M = 1$. Thus, (2) reduces to

$$r(\mathbf{X}) = \underset{k}{\operatorname{argmax}} \Big\{ \max_{\boldsymbol{\vartheta}} \{ p(\boldsymbol{\vartheta})\ p(k)\ p(\mathbf{X}_{I_0} | \mu_0)\ p(\mathbf{X}_{I_1} | \boldsymbol{\mu}_k(\boldsymbol{\vartheta})) \} \Big\}. \tag{3}$$

We consider 2D-rotation, scaling with fixed aspect ratio, and translation as transformations. The priors $p(\boldsymbol{\vartheta})$ and $p(k)$ are assumed uniform. The object density $p(\mathbf{X}|\boldsymbol{\mu}_k)$ is modeled using Gaussian kernel densities or Gaussian mixture densities. The use of mixture models allows the implicit modeling of further transformations by mapping them to different densities if they are observed in the training data. The part of the image that is not assigned to any object is assigned to the class background. In the experiments, the set of background pixels is modeled by a univariate distribution on the pixel level, where individual pixel values are assumed to be statistically independent. I.e. we assume for the background model $p(\mathbf{X}|\mu_0) = \prod_{x \in \mathbf{X}} p(x|\mu_0)$. The local density $p(x|\mu_0)$ is chosen among univariate Gaussian, uniform distribution, or empirical histograms with different numbers of bins. Note that the correct normalization of the distributions is important because of the changing amount of pixels that are explained for different transformation parameters $\boldsymbol{\vartheta}$. One example partitioning is shown in Fig. 1.

Decision Making. To illustrate the search or decision problem arising from the decision rule (3), we fix the hypothesized class k and assume the maximizing

transformation parameters $\hat{\boldsymbol{\vartheta}}$ are to be determined. E.g. considering Gaussian densities $p(\mathbf{X}|\boldsymbol{\mu}_k) = \mathcal{N}(\mathbf{X}|\boldsymbol{\mu}_k, \sigma_1^2\mathbf{I})$ for the objects and $p(x|\mu_0) = \mathcal{N}(x|\mu_0, \sigma_0^2)$ for the background leads to the search

$$\begin{aligned}\hat{\boldsymbol{\vartheta}} &= \operatorname*{argmax}_{\boldsymbol{\vartheta}} \{p(\boldsymbol{\vartheta})\ p(k)\ p(\mathbf{X}_{I_0}|\mu_0)\ p(\mathbf{X}_{I_1}|\boldsymbol{\mu}_k(\boldsymbol{\vartheta}))\} \qquad (4)\\ &= \operatorname*{argmin}_{\boldsymbol{\vartheta}} \big\{ -\log p(\boldsymbol{\vartheta}) - \log p(k) + \frac{1}{2}|I_0| \log(2\pi\sigma_0^2) + \frac{1}{2\sigma_0^2} \sum_{x \in \mathbf{X}_{I_0}} (x-\mu_0)^2 \\ &\qquad + \frac{1}{2}|S_1| \log(2\pi\sigma_1^2) + \frac{1}{2\sigma_1^2} \|\mathbf{X}_{I_1} - \boldsymbol{\mu}_k(\boldsymbol{\vartheta})\|^2 \big\}\end{aligned}$$

The large number of parameter settings $\boldsymbol{\vartheta}$ makes the search for the maximizing arguments a complex problem. Optimization strategies should be considered:

- The Euclidean distances $\|\mathbf{X}_{I_1} - \boldsymbol{\mu}_k(\boldsymbol{\vartheta})\|$ for all translations can be efficiently calculated using the fast Fourier transform reducing the computation effort for this term in the order of $\log |\mathbf{X}| \ / \ |\boldsymbol{\mu}_k(\boldsymbol{\vartheta})|$.
- The sums of squares $\sum_{x \in \mathbf{X}_{I_0}} (x-\mu_0)^2$ for all translations can be efficiently computed using precomputed sums of squares. This reduces the effort for this term in the order of $|\boldsymbol{\mu}_k(\boldsymbol{\vartheta})|^{-1}$.
- The search space can be reduced by limiting the number of hypothesized transformations or by restricting the regions I_1 to square regions.
- A significant speedup can be gained by pruning the search space using the results of a complete search in a down-scaled version of the scene.

Training. Algorithms for single object recognition cannot be used to determine the model parameters without given segmentation. The following training algorithm is based on an expectation-maximization (EM) scheme, where the hidden variables are the parameters $\boldsymbol{\vartheta}$ for each object in each training scene:

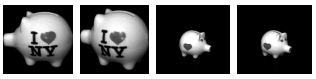
1. initialize model parameters
2. search maximizing transformation parameters $\boldsymbol{\vartheta}$ in each scene using (4)
3. re-estimate model parameters (e.g. EM algorithm for mixtures)
4. repeat from 2 until convergence

For the training we assume exactly one object to be present in each image. Furthermore, objects are assumed to lie within a square region. The initial model parameters can be based on a constant graylevel estimated from a histogram of the training data or a small set of manually segmented objects. The latter approach facilitates convergence and still leads to a high reduction of manual preprocessing. The hypothesized transformations are translation, scaling with fixed aspect ratio and 2D-rotation.

3 Databases and Results

To evaluate the quality of an image classification approach it is important to compare the obtained results to those of other methods on the same data. One of the drawbacks in the research within this field is that there exists no widely used

Table 1. Description of databases COIL-20 and ERLANGEN with error rates (ER).

name	ERLANGEN	COIL-20
# classes	5	20
# training images	90	720
# test images	85	180
example images		
other methods (ER [%])	[6] 0.0	[2] 0.0
holistic model (ER [%])	0.0	0.0

benchmark database for object-based image recognition or object training. Many groups use their own non-public data which makes it impossible to compare results. A number of databases exist for different purposes, as e.g. face recognition or handwritten digit recognition, or the used databases contain unspecific images on which the results are judged qualitatively by a human observer.

The website `http://www-2.cs.cmu.edu/~cil/v-images.html` lists many databases used in computer vision research, out of which none is suitable for this task. An exception is a database of images collected by the authors of [6], although error rates of 0% can be achieved, making a comparison difficult. Due to this absence of a standard benchmark we created a database for object-based scene analysis based on the well known Columbia Object Image Library (COIL) and a set of real-world backgrounds. This database named COIL-RWTH is publicly available upon request and results are presented in Section 3.3. Tables 1 and 3 show an overview of the databases used in this work.

3.1 ERLANGEN

Database. In [6] the authors used two databases of images containing five different objects, all images of size 256×256. The first of the databases contains images taken with one illumination while in the second case the objects are illuminated with two light sources. Each of the training sets contains 18 images per object taken at different 2D rotation angles on a homogeneous background. Another 17 images per object at rotation angles not occurring in the training set are in the test sets. For each database, three different test sets exist, one with heterogeneous background, and two with two different levels of occlusion. Note that background and occlusions were added to the images artificially. Note also that the background is identical in all of the images and it does not occur in the training images as background (although one image containing only the background exists). The background resolution differs from that of the object images, which might be advantageous when using features based on Gabor filters.

Results. We used the test set with heterogeneous background from the first database and the corresponding training set. In [6] a recognition error rate of 0% is reported. The same error rate was achieved using the proposed holistic model with rectangular prototype models.

3.2 COIL-20

Database. The Columbia Object Image Library (COIL-20) [5] contains 72 graylevel images for each of a set of 20 different objects, taken at intervals of five degrees 3D-rotation. To strictly separate train and test images, we use the odd angles of the 'processed' corpus (size 128×128) for training and the even angles of the 'unprocessed' corpus (size 448×416) for testing. The two corpora differ in the lighting conditions (because of the processing) and the size of the object in the image (cp. Table 1). This procedure ensures at least 5 degrees difference in 3D position and poses the additional difficulty of differing light conditions. Other authors use a splitting of the 'processed' corpus into train and test, but in this case even a Euclidean nearest neighbor classifier leads to a 0% error rate.

Results. On the original COIL-20 database, the holistic approach achieves a 0% error rate without further tuning than using a Gaussian background model with mean zero and low variance. This result seems not surprising, as the images are shown on a homogeneous black background. But as the training and test images appear at different lighting conditions and on different scales, a nearest neighbor classifier is not sufficient for completely correct classification and it is necessary to extend it with elaborate techniques to achieve a 0% error rate [2].

3.3 COIL-RWTH

Database. As the COIL-20 database only contains images with homogeneous black background, segmentation of the object from the background is a feasible approach to classification. On the other hand, for real-world images segmentation poses a serious problem. (Although many application areas exist, where a homogeneous or static background can be assumed and existing methods provide acceptable solutions.) Therefore, a new dataset was created based on the objects from the COIL-20 database and a set of new background images. The goal was to create tasks of increasing difficulty to extend the COIL-20 task that can be solved perfectly by existing methods. Each test image carries information about the used transformation parameters for the object images, allowing to separate the effects of different transformations.

We created two corpora that differ in the background used: The COIL-RWTH-1 corpus contains objects placed on a homogeneous black background, whereas the COIL-RWTH-2 corpus contains the objects in front of inhomogeneous real-world background images that were kept separate for training and test images and vary in resolution. The two training and test sets are based on the COIL-20 sets as described above. The training images are of size 192×192 and the size of the test images is 448×336. In all sets, we applied the following uniformly distributed random transformations to the object images: translation, 360 degree 2D-rotation, and 60–100% scaling with fixed aspect ratio.

Results. To investigate the effect of different background models, we tested univariate Gaussian densities, uniform distributions, and histograms with varying numbers of bins. In about 70% of the evaluated experiments, the univariate Gaussian densities performed best among these models [4]. In the following we therefore only discuss results obtained with this background model.

Table 2. Training results for Gaussian single densities on COIL-RWTH-2 with fixed 3D-rotation angle shown for one of the objects.

	rotation known		scaling known		no information	
initial mean of Gaussian density						
resulting mean of Gaussian density						

To observe the effect of known transformation parameters on the proposed training, we trained a Gaussian single density on all images with a fixed 3D-rotation angle of COIL-RWTH-2. The resulting mean images are shown in Table 2. It can be observed that the algorithm finds visually important parts of the object searched for. The exact appearance of the mean image differs strongly depending on the used initialization and the information supplied to the training algorithm. To evaluate the proposed training algorithm further, we trained Gaussian mixture densities on COIL-RWTH-1 and used these models to classify the original COIL-20 dataset. This resulted in 7.8% error rate. Note that the mixture density now models the different 3D-rotation angles of the objects. If the correct 2D-rotation of the object is supplied to the training algorithm, this error rate can be reduced to 4.4%. To separate the effect of unknown rotation from the other unknown parameters, in the following we only present results, in which the 2D-rotation of the objects in the images is known to the classifier.

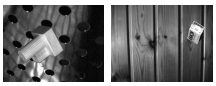
We evaluated the classification accuracy of the complete setup on the COIL-RWTH databases in three scenarios. The results are shown in Table 3. As no other results are available, we used a conventional kernel density classifier for comparison. This classifier was supplied with the same information and an object position compensation was implemented using the center of gravity of the images. The results show that the holistic model performs with acceptable error rates for homogeneous background. Recall that scale changes are handled automatically and segmentation is performed implicitly in the model. The high error rates of the kernel density classifier can be explained by the fact that it cannot cope with scale changes. This also explains the improving error rate for the COIL-RWTH-1 test data when switching from the COIL-20 to the COIL-RWTH-1 training data, because the latter already includes variations in scale.

The error rates for the inhomogeneous background are clearly inacceptable. The failure of the algorithm here is based on the coincidence of two problems: 1. Automatic object training with unknown segmentation and variable background is very difficult. The resulting mean vectors show strong blur due to the changing background but capture some characteristic information, which is not enough to achieve lower error rates. 2. Detection of objects of variable scale and position in large inhomogeneous images based on an incomplete object model of graylevels and backgrounds not seen in training is possible only in few cases.

4 Conclusion

We presented a holistic statistical model for appearance-based training and recognition of objects in complex scenes. Experiments on two existing databases

Table 3. Error rates for the COIL-RWTH database (20 classes, 180 test images each).

training data	COIL-20 720 images 	COIL-RWTH-1 5760 images 	COIL-RWTH-2 5760 images 
test data	COIL-RWTH-1 	COIL-RWTH-1 	COIL-RWTH-2 
kernel dens. (ER[%])	38.9	27.2	95.0
holistic (ER[%])	1.1	7.8	92.8

show the algorithm to be competitive with other known approaches. A third database with a higher level of difficulty that can be used by other researchers was introduced. The gained results underline the difficulty of training and recognition in the presence of inhomogeneous background. The fact that the presented method achieves 0% error rates on two databases used in the literature, but fails on a database of images with highly misleading background shows that the databases on which 0% error rates can be reported are by far not representative for the complexity of the general object-based scene analysis problem.

Most improvement to the presented method can be expected from the inclusion of more descriptive features than only grayvalues, like e.g. wavelet features or local representations of image parts. Furthermore, local variations of the objects may be modeled using tangent distance or appropriate distortion models.

Acknowledgements. We would like to thank the members of the Chair for Pattern Recognition, Department of Computer Science, Friedrich Alexander University of Erlangen-Nürnberg for providing their database, and the members of the Department of Computer Science, Columbia University, New York for sharing their data openly.

References

1. B.J. Frey, N. Jojic: Transformation-invariant clustering using the EM algorithm. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(1):1–17, January 2003.
2. D. Keysers, J. Dahmen, H. Ney, M.O. Güld: A Statistical Framework for Multi-Object Recognition. In *Informatiktage 2001 der Gesellschaft für Informatik*, Konradin Verlag, Bad Schussenried, Germany, pp. 73–76, October 2001.
3. D. Keysers, J. Dahmen, H. Ney, B. Wein, T. Lehmann: Statistical Framework for Model-based Image Retrieval in Medical Applications. *J. Electronic Imaging*, 12(1):59–68, January 2003.
4. M. Motter: Statistische Modellierung von Bildobjekten für die Bilderkennung. Diploma thesis, Chair of Computer Science VI, RWTH Aachen University of Technology, Aachen, Germany, December 2001.
5. H. Murase, S. Nayar: Visual Learning and Recognition of 3-D Objects from Appearance. *Int. J. Computer Vision*, 14(1):5–24, January 1995.
6. M. Reinhold, D. Paulus, H. Niemann: Appearance-Based Statistical Object Recognition by Heterogeneous Background and Occlusions. In *Pattern recognition. 23rd DAGM Symposium*, LNCS 2191, Munich, Germany, pp. 254–261, September 2001.
7. A.W.M. Smeulders, M. Worring, S. Santint, A. Gupta, R. Jain: Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1349–1380, December 2000.