Michael Reinhold, Dietrich Paulus, Heinrich Niemann

Appearance-Based Statistical Object Recognition by Heterogeneous Background and Occlusions

appeared in:
Pattern Recognition, 23rd DAGM Symposium,
12.-14. September 2001, München,
Springer, Lecture Notes in Computer Science 2191, p. 254-261

Appearance-Based Statistical Object Recognition by Heterogeneous Background and Occlusions

Michael Reinhold *, Dietrich Paulus, and Heinrich Niemann

Chair for Pattern Recognition, University Erlangen-Nürnberg
Martensstr. 3, 91058 Erlangen, Germany
reinhold@informatik.uni-erlangen.de,
http://www5.informatik.uni-erlangen.de

Abstract In this paper we present a new approach for the localization and classification of 2-D objects that are situated in heterogeneous background or are partially occluded. We use an appearance-based approach and model the local features derived from wavelet multiresolution analysis by statistical density functions. In addition to the object model we define a new model for the background and a function that assigns the single feature vectors either to the object or to the background. Here, the background is modelled as uniform distribution, therefore we need for all possible backgrounds only one density function. Experimental results show that this model is well suited for this recognition task.
Keywords: object recognition, statistical modelling, background model

1 Introduction

There are several approaches for object recognition. Since the approaches that use the results of a segmentation process like lines or vertices suffer from segmentation errors and the loss of information, we use an appearance-based approach; i. e. the image data, e. g. the pixel intensities, are used directly without a previous segmentation process. One appearance-based approach is the "eigenspace" introduced by [5] that use principal component analysis and encode the data in so-called "eigen-images". Other authors apply histograms; the most well-known approach are the "multidimensional receptive field histograms" of [8] which contain the results of local filtering, e. g. by Gaussian derivative filters. [2] use a statistical approach for the recognition and model the features by Gaussian mixtures. We use and extend the statistical model of [6]: local feature vectors are calculated by the coefficients of the multiresolution analysis using Johnston-Wavelets. The object features are modelled statistically as normal distributions by parametric density functions. As we will show in the experiments this approach is very insensitive to changes in the illumination.

For the recognition in real environments very often the objects reside not in the homogeneous background of the training, but in cluttered background and some parts of the objects are occluded. For this purpose [4], who use the eigenspace approach, try to find n object features in the image that are neither affected by the background nor

* This work was funded by the German Science Foundation (DFG) Graduate Research Center 3-D Image Analysis and Synthesis. Only the authors are responsible for the content.

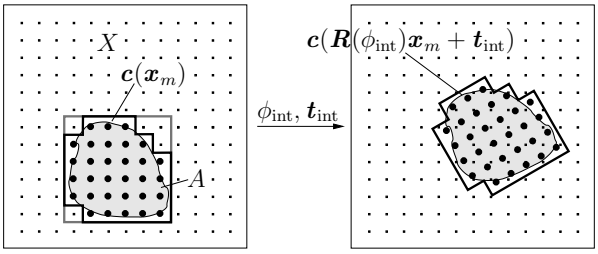

Figure 1. *left:* Image covered by a grid, the object is enclosed by a tight boundary (black line); the respective old rectangular box is plotted in gray, because of its form it enclose not only object features but also background features. *right:* For movements of the object inside the image plane the object grid is transformed with the same internal rotation ϕ_{int} and internal translation $\boldsymbol{t}_{\mathrm{int}}$ as the object

occluded. For this reason they generate pose hypotheses and select the n best fitting features. In contrast other authors explicitly model the background and assign the features to the object or to the background. For example, [3], who use vertices as features, assume a uniform distribution for the position of the background features and model a probabilistic weighted assignment. Whereas in [1] the features are assigned neither to the object nor to the background. Since medical radiographs were classified, the background model has a constant gray value of zero. Also for our approach a background model was presented in [6] and [7]: the known background was trained as Gaussian distribution and a weighted assignment was applied. We propose a new model for this approach: the background is modelled as uniform distribution over the possible values of the feature vectors, therefore we are independent of the current background. We define a assignment function that assigns each of the local feature vectors either to the object or to the background, depending whether the calculated value of the object density or of the background density is higher for this local feature vector. The recognition is performed hierarchically by a maximum likelihood estimation, whereby an accelerated algorithm is used for the global search.

In the following section we shortly outline the object model for homogeneous background and in section 3 we describe our new model for heterogeneous background and occlusions. In section 4 the experiments and the results are presented. Finally we end with a summary of the results and the future work in section 5.

2 Object Model for Homogenous Background

A grid with the sampling resolution r_s, whereby s is the index for the scale, is laid over the square image $\boldsymbol{f}$ as one can see in figure 1. These grid locations will be summarized in the following as $X_s = \{\boldsymbol{x}_{m,s}\}_{m=0,\ldots,M-1}, \boldsymbol{x}_{m,s} \in \mathbb{R}^2$. On each grid-point a local feature vector $\boldsymbol{c}_s(\boldsymbol{x}_{m,s})$ with two components is calculated by the coefficients of the multiresolution analysis using discrete Johnston 8-TAP wavelets and is interpreted as random variable for the statistical model. Thereby the randomness among others is the

noise of the image sampling process and changes in the lighting conditions. To simplify the notation, the index s is omitted in the following.

For the object model a close boundary is laid around the object. In [6] only a rectangular box was implemented and it was positioned manually. In contrast now the form of the boundary is arbitrary so that it enclose the object much better (see left image of figure 1). Besides it is calculated automatically during the training, wherefore only one image of the object in front of a dark background is necessary. During the recognition this trained form of the bounded region A is used. We assume that the feature vectors $\boldsymbol{c}_{A,m}$ inside the bounded region $A \subset X$ belong to the object, and are statistically independent from the feature vectors $\boldsymbol{c}_{X \setminus A,k}$ outside the bounded region $X \setminus A$. Therefore, for the object model we only need to consider the (object) feature vectors $\boldsymbol{c}_{A,m}$, concatenated written as vector $\boldsymbol{c}_A$. Now, the object can be described by the density function $p(\boldsymbol{c}_A|\boldsymbol{B}, \phi, \boldsymbol{t})$ depending on the learned model parameter set $\boldsymbol{B}$, the rotation $\phi = \phi_{\mathrm{int}}$ and the translation $\boldsymbol{t} = \boldsymbol{t}_{\mathrm{int}}$.

Further we assume that the features are normally distributed. In [6] a statistical dependency between adjacent feature vectors in a row was modelled, but for arbitrary objects the results is worse than for statistical independency. Therefore we assume that the single feature vectors are statistically independent from each other. So the density functions can be written as

$$p(\boldsymbol{c}_A|\boldsymbol{B}, \phi, \boldsymbol{t}) = \prod_{\boldsymbol{x}_m \in A} p(\boldsymbol{c}_{A,m}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m, \phi, \boldsymbol{t}) \quad , \tag{1}$$

whereby $\boldsymbol{\mu}_m$ is the mean vector and $\boldsymbol{\Sigma}_m$ the covariance matrix of the m-th feature vector. Because of the statistically independence $\boldsymbol{\Sigma}_m$ is a diagonal matrix.

For the localization we perform a maximum likelihood estimation over all possible rotations $\phi = \phi_{\mathrm{int}}$ and transformations $\boldsymbol{t} = \boldsymbol{t}_{\mathrm{int}}$:

$$(\phi, \boldsymbol{t}) = \underset{(\phi, \boldsymbol{t})}{\operatorname{argmax}}\, p(\boldsymbol{c}_A|\boldsymbol{B}_\kappa, \phi, \boldsymbol{t}) \quad , \tag{2}$$

and for the classification an additional maximum likelihood estimation over all possible classes κ:

$$(k, \phi, \boldsymbol{t}) = \underset{\kappa}{\operatorname{argmax}} \underset{(\phi, \boldsymbol{t})}{\operatorname{argmax}}\, p(\boldsymbol{c}_A|\boldsymbol{B}_\kappa, \phi, \boldsymbol{t}) \quad . \tag{3}$$

For accelerating the localization, first a rough localization is conducted on a rough resolution, followed by a refinement on a finer resolution. For further details see [6].

3 Model for Heterogenous Background and Occlusions

For occlusions the assumption that all the feature vectors inside the region A belong to the object is wrong. Besides for heterogeneous background the features vectors at the border of the object that cover not only the object but also partially the background are modified. Therefore [4] try to find n of the totally N object features that are not affected by the heterogeneous background and the occlusion [4]. But for this approach there is a risk to confuse similarly looking objects - like the two matchboxes in figure 3 (below).

For this purpose, we consider all local feature vectors $\boldsymbol{c}_{A,m}$ in the bounded region A and define a background model and a assignment function $\boldsymbol{\zeta} \in \{0,1\}^N$. Whereby the m-th component ζ_m of $\boldsymbol{\zeta}$ assigns the local feature vector $\boldsymbol{c}_{A,m}$ to the object ($\zeta_m = 1$) or to the background ($\zeta_m = 0$). So the density function

$$p(\boldsymbol{c}_A|\boldsymbol{B},\phi,\boldsymbol{t}) = \sum_{\boldsymbol{\zeta}} p(\boldsymbol{c}_A,\boldsymbol{\zeta}|\boldsymbol{B},\phi,\boldsymbol{t}) \tag{4}$$

also includes the assignment function $\boldsymbol{\zeta}$ and becomes a mixture density. Now $\boldsymbol{B}$ includes the learned parameters $\boldsymbol{B}_1$ of the object as well the learned parameters $\boldsymbol{B}_0$ of the background.

In [6] and [7] the background has to been known already during the training and was trained as a normal distribution, i. e. for each different background an own background density has to been trained. Besides statistically dependencies between the feature vectors in a row were modelled. For some experiments also for the assignment function a row-dependency was modelled [7], whereas for other experiments statistically independence of the assignments was supposed [6]. Thereby weighted assignments were modelled, i. e. a local feature vector $\boldsymbol{c}_{A,m}$ belongs with a probability $p(\zeta_m = 1)$ to the object and $p(\zeta_m = 0)$ to the background. This leads to a very complex model.

To be independent from the current background and handle every possible background by only one background density, we model the background as a uniform distribution over the possible values of the feature vectors. So nothing has to be known about the background a priori; the background density depends only on chosen feature set, e. g. the wavelet type for filtering. Additionally, it is identical for all feature vectors and therefore it is independent from the transformations $\boldsymbol{t}_{\text{int}}$ and ϕ_{int}. As in section 2 we assume statistically independence of the features and also of the assignments, so the density function in (4) can be transformed to

$$\begin{aligned} p(\boldsymbol{c}_A|\boldsymbol{B},\phi,\boldsymbol{t}) &= \sum_{\boldsymbol{\zeta}} \prod_{\boldsymbol{x}_m \in A} p(\boldsymbol{c}_{A,m},\zeta_m|\boldsymbol{B},\phi,\boldsymbol{t}) \\ &= \prod_{\boldsymbol{x}_m \in A} \sum_{\zeta_m} p(\boldsymbol{c}_{A,m},\zeta_m|\boldsymbol{B},\phi,\boldsymbol{t}) \\ &= \prod_{\boldsymbol{x}_m \in A} \sum_{\zeta_m} p(\zeta_m)\, p(\boldsymbol{c}_{A,m}|\zeta_m,\boldsymbol{B},\phi,\boldsymbol{t}) \quad . \end{aligned} \tag{5}$$

This is a much simpler expression than (4): now we no longer have a marginalization about all possible assignments $\boldsymbol{\zeta}$ for all features $\boldsymbol{c}_A$, but for each single feature vector $\boldsymbol{c}_{A,m}$ a marginalization about the single assignments ζ_m.

The assignment ζ_m is a hidden information. For example we do not know, how much of an object is occluded. So we set the a priori probability that a local feature vector $\boldsymbol{c}_{A,m}$ belongs to the object or to the background as equal. Further we model ζ_m as a (0,1)-decision, i. e. a feature vector $\boldsymbol{c}_{A,m}$ belongs either to the object or to the background. The decision is taken during the localization process. Thereby $\boldsymbol{\zeta}$ is chosen so that the density value $p(\boldsymbol{c}_A|\boldsymbol{B},\phi,\boldsymbol{t})$ is maximized. That is the density value for each

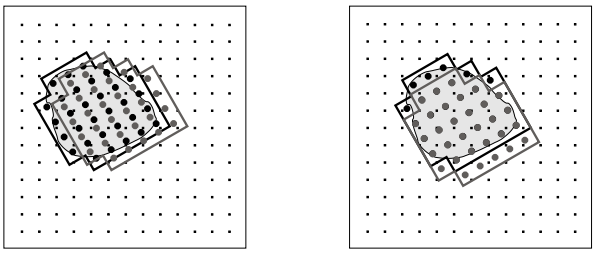

Figure 2. *left:* for each possible internal transformation all the feature vectors have to been interpolated, *right:* for the same rotation only another translation the most feature vectors can be reused

feature vector $\boldsymbol{c}_{A,m}$ (see (5))

$$\sum_{\zeta_m} p(\zeta_m)\; p(\boldsymbol{c}_{A,m}|\zeta_m, \boldsymbol{B}, \phi, \boldsymbol{t}) \tag{6}$$

has to be maximized. This can be done by the assignment ζ_m

$$\zeta_m = \underset{\zeta_m}{\operatorname{argmax}}(p(\boldsymbol{c}_{A,m}|\zeta_m = 1, \boldsymbol{B}_1, \phi, \boldsymbol{t}), p(\boldsymbol{c}_{A,m}|\zeta_m = 0, \boldsymbol{B}_0)) \quad , \tag{7}$$

and setting the probability $p(\zeta_m)$ for the respective assignment to one, the probability for the other assignment to zero.

For example, for the Johnston wavelets used in the experiments of section 4 the object density $p(\boldsymbol{c}_{A,m}|\zeta_m = 1, \boldsymbol{B}_1, \phi, \boldsymbol{t})$ for a not occluded feature vector $\boldsymbol{c}_{A,m}$ lays typically between e^{-1} and e^{1}. For occlusion it becomes very low: between e^{-100} and e^{-10}. In this case the feature vector $\boldsymbol{c}_{A,m}$ is assigned to the background and the background density $p(\boldsymbol{c}_{A,m}|\zeta_m = 0, \boldsymbol{B}_0) = e^{-3.5}$ (that is the value for the used Johnston wavelets) is chosen for this feature vector $\boldsymbol{c}_{A,m}$.

For the localization and classification a maximum likelihood estimation is performed as described in (2) and (3). We also tested a heuristic measurement in section 4 for the classification. The single objects differ in their size, i. e. also in the number N of their local object feature vectors $\boldsymbol{c}_{A,m}$ inside the bounded region A. Since non fitting object features are assigned to the background and the background density value is used, there are some misclassification caused by the different size of the objects. Therefore we normalize the density before the maximum likelihood estimation by the number N of local object feature vectors $\boldsymbol{c}_{A,m}$:

$$(k, \phi, \boldsymbol{t}) = \underset{\kappa}{\operatorname{argmax}} \underset{(\phi,\boldsymbol{t})}{\operatorname{argmax}} \sqrt[N]{p(\boldsymbol{c}_A|\boldsymbol{B}_\kappa, \phi, \boldsymbol{t})} \quad . \tag{8}$$

Because of the statistical independence (see (5)) the expression in (8) is the geometric mean of the density values $p(\boldsymbol{c}_{A,m}|\zeta_m, \boldsymbol{B}, \phi, \boldsymbol{t})$ of all feature vectors $\boldsymbol{c}_{A,m}$ with the respective assignments ζ_m.

Further we speed up the first step of the localization process that starts with a global search over all possible internal rotations ϕ_{int} and translations $\boldsymbol{t}_{\text{int}}$ on a rough resolu-

Figure 3. The 5 objects in the different environments: box on the training background, matchbox 1 on the black background, matchbox 2 on the heterogenous background, car 1 with 25% black occlusion, car 2 with 50% heterogeneous occlusion

tion. Although we evaluate the density function only at discrete points of this transformation space, we have to calculate the density value for 225 possible internal translations $\boldsymbol{t}_{\mathrm{int}}$ each with 36 possible internal rotations ϕ_{int}. For each of the altogether 8100 possible transformations we have to interpolate about the 80 feature vectors as one can see in the left image of figure 2.

Since the interpolation is computationally expensive, we change this algorithm. We interpolate the required area of the grid for each rotation ϕ_{int} only once and then translate the object grid according to the rotated coordinates axes in steps respective to the resolution r_s. As visible in the right image of figure 2 each interpolated feature vector can be used for many transformations.

4 Experiments and Results

For the experiments we used the five objects in figure 3. The images were 256 pixels in square. For the training, 18 images of each object in different poses with the same illumination were taken, the background was nearly homogeneous with a pixel intensity about 60. We took 17 further images of each object in other positions for the tests. For the experiments with heterogeneous background we cut out the objects and pasted them in an absolute black background as well as in front of a mouse pad (see figure 3). For the occlusions we blacked out 25% respectively 50% of the object in the test images, as well as we covered the objects in the image with a part of the mouse pad. It has

Table 1. Error rate in percent for only object density no background model, the old background model [6], the new background model without and with normalization (see eq. 8); for each model $3*170$ localization and classification experiments are performed, *left:* error rate for the localization, *middle:* for the classification, *right:* average computation time for one localization on a Pentium III 800 MHz

	localization			classification			
one illumination	het. back.	25% occl.	50% occl.	het. back.	25% occl.	50% occl.	time
only object dens.	22,9%	69,4%	82,4%	25,3%	62,4%	70,6%	0,8 s
background m. [6]	6,5%	24,1%	51,7%	20,1%	21,2%	50,0%	6,5 s
new background m.	0,0%	0,0%	7,1%	0,0%	0,0%	4,7%	1,3 s
the same with norm.				0,0%	0,0%	2,3%	1,3 s

to be mentioned that a absolute black background or occlusion is a big difference to the training, because the first component of the local feature vectors is the logarithmic low-pass value of the wavelet analysis. So we got for each model 170 localization and 170 classification experiments for heterogeneous background, 25% occlusion and 50% occlusion. For the background model [6] two background classes were trained and used: the absolute black background and the mouse pad.

For the recognition we used for the rough localization a resolution of $r_s = 8$ pixels and for the refinement a resolution of $r_s = 4$ pixels. The objects were searched in the whole image for all internal rotations and translations. Thereby a localization is counted as wrong if the rotation error is bigger than 7° respectively the translation error bigger than 7 Pixels.

As one can see in table 1 the simple object model for homogeneous background (section 2) could handle heterogenous background, but failed very often for occlusion. The old background model [6] is better than the simple object model, but it is very slow and for 50% occlusion the error rate was about 50%. Whereas the new background model is much better and faster: for heterogeneous background and 25% occlusion there were no errors, and even for 50% occlusion the error rate was small. Further, by the normalization the classification error rate for 50% occlusion could be reduced from 4,7% to 2,3%. Additionally, the new model is five times faster than the old background model.

For testing the robustness of this approach we also performed experiments with two different illuminations. We trained each object with 9 images taken with illumination 1 and 9 images taken with illumination 2, where one of the three spotlights was switched off. Also for the test images we used these two illuminations. For the new background model the localization and classification error rate for heterogenous background and 25% occlusion was still very small, only for 50% occlusion it increased. But it could be reduced to 4,8% by the normalization.

5 Conclusions and Outlook

In this paper we presented a new efficient background model for object recognition. The background is modelled as uniform distribution and therefore independent from

Table 2. Error rates in percent for two illuminations: for only object density no background model, the old background model [6], the new background model without and with normalization (see eq. 8); for each model $3*170$ localization and classification experiments are performed, *left:* error rate for the localization; *right:* for the classification, the average computation time is nearly the same as in table (1)

	localization			classification		
two illuminations	het. back.	25% occl.	50% occl.	het. back.	25% occl.	50% occl.
only object dens.	11,4%	60,0%	77,7%	26,4%	50,0%	70,0%
background m. [6]	6,8%	32,4%	54,7%	7,7%	23,9%	61,2%
new background m.	0,0%	3,0%	24,2%	0,0%	4,7%	18,9%
the same with norm.				0,0%	0,0%	4,8%

the current used background, i. e. all possible backgrounds can be handled by only one background density function. We defined a assignment function that assigns each local feature vector either to the object or to the background. With this model we improved the recognition rate to nearly 100% for heterogeneous background and 25% occlusion and to nearly 95% for occlusion of 50%, even if two different lighting conditions are used.

In the future we will extend the background model to 3-D objects. For this purpose we have to model the so far fix size of the bounded region as a function of the out of image plane transformations. This is necessary because the appearance and the size of the objects vary with these external transformations. In addition the assignment function is a good basis for multi object recognition.

References

1. J. Dahmen, D. Keysers, M. Motter, H. Ney, T. Lehmann, and B. Wein. An automatic approach to invariant radiograph classification. In H. Handels, A. Horsch, T. Lehmann, and H.-P. Meinzer, editors, *Bildverarbeitung für die Medizin 2001*, pages 337–341, Lübeck, March 2001. Springer Verlag, Berlin.
2. J. Dahmen, R. Schlüter, and H. Ney. Discriminative training of gaussian mixtures for image object recognition. In W. Förstner, J. Buhmann, A. Faber, and P. Faber, editors, *21. DAGM Symposium Mustererkennung*, pages 205–212, Bonn, September 1999. Springer Verlag, Berlin.
3. J. Hornegger. *Statistische Modellierung, Klassifikation und Lokalisation von Objekten*. Shaker Verlag, Aachen, 1996.
4. A. Leonardis and H. Bischof. Dealing with occlusions in the eigenspace approach. In *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 453–458, San Francisco, 1996.
5. H. Murase and S. K. Nayar. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.
6. J. Pösl. *Erscheinungsbasierte statistische Objekterkennung*. Shaker Verlag, Aachen, 1998.
7. J. Pösl and H. Niemann. Object localization with mixture densities of wavelet features. In *International Wavelets Conference*, Tanger, Marokko, April 1998. INRIA.
8. B. Schiele and J. Crowley. Object recognition using multidimensional receptive field histograms. In *Fourth European Conference on Computer Vision (ECCV)*, pages 610–619, Cambridge, UK, April 1996. Springer, Heidelberg.