Statistical Classifiers in Computer Vision

Joachim Hornegger, Dietrich Paulus, Heinrich Niemann
Lehrstuhl für Mustererkennung (Informatik 5)

hornegger@informatik.uni-erlangen.de
paulus@informatik.uni-erlangen.de
niemann@informatik.uni-erlangen.de

Reprint of the proceedings

Data Highways and Information Flooding, a Challenge for Classification and Data Analysis

Potsdam, Germany, 12. – 14. March 1997

Joachim Hornegger, Dietrich Paulus, Heinrich Niemann: *Statistical Classifiers in Computer Vision*, in I. Balderjahn, R. Mathar, M. Schader (Eds.): *Data Highways and Information Flooding, a Challenge for Classification and Data Analysis*, Springer, Berlin, to appear 1997.

Contents

Statistical Classifiers in Computer Vision

J. Hornegger, D. Paulus, H. Niemann[1]

Lehrstuhl für Mustererkennung (Informatik 5),
Universität Erlangen, Martensstraße 3, D-91058 Erlangen, Germany

Abstract: This paper introduces a unified Bayesian approach to 3–D computer vision using segmented image features. The theoretical part summarizes the basic requirements of statistical object recognition systems. Non–standard types of models are introduced using parametric probability density functions, which allow the implementation of Bayesian classifiers for object recognition purposes. The importance of model densities is demonstrated by concrete examples. Normally distributed features are used for automatic learning, localization, and classification. The contribution concludes with the experimental evaluation of the presented theoretical approach.

1 Introduction

Classification in computer vision is commonly dominated by geometrical, model–based approaches (Faugeras (1993)). Heuristics for many algorithms in image processing restricted to the given problem domain and motivated by associated applications are reported in the literature. Herein, model–based image analysis provides the scientific framework for matching algorithms and for understanding the information process. The comprehensive goal is to describe the intrinsic character of images in a symbolic or parametric manner.

Bayesian methods have provided solutions to various classical problems in pattern recognition. Especially the progress in the field of speech processing is substantially based on the application of statistical methods. The general use of Bayesian classifiers is motivated by several aspects: they show optimality in a decision theoretic sense under a 0–1 cost function (Duda and Hart (1973)). Furthermore, statistical methods can deal with uncertainty in a natural manner, have a well elaborated mathematical theory, and provide a unified framework within which many different tasks can be considered. For that reason, we favor model–based computer vision algorithms which apply statistical discriminants or, at least, close approximations of Bayesian classifiers.

In this paper, we present a probabilistic framework for 3–D vision: statistical methods for object modeling, algorithms for the automatic estimation of model parameters — even in the presence of incomplete and disturbed training data —, classification rules, and localization methods for 3–D objects using 2–D views. The introduced model densities show several degrees of freedom, and standard hidden Markov models or mixtures of densities can be derived by specialization. The experiments prove that the classification and pose estimation task for 3–D objects using real image data can be treated statistically.

[1]The authors wish to thank the German Research Foundation (DFG), who partially funded the work reported here under grant SFB 182.

A general discussion of Bayesian image analysis (section 2) is followed by a statistical description of objects and their appearance in scenes (section 3). The object recognition and localization problem is formalized (section 4), and experimental results for these problems are given (section 5).

2 Bayesian Image Analysis

There exists a wide range of model–based methods for computer vision. Model–based statistical algorithms, in general, require the stages model selection, sampling, parameter estimation, and goodness–of–fit. The main difference between standard geometrical techniques and probabilistic modeling schemes is due to the fact that Bayesian image analysis methods make use of statistical models to incorporate both, general and object specific prior knowledge. The object recognition problem is understood as the assignment of a subset of observed image features to a pattern class Ω_κ $(1 \leq \kappa \leq K)$, which characterize one object or a set of objects. Statistical classifiers known from pattern recognition theory require feature vectors $\boldsymbol{c}$ of fixed dimensions and a probabilistic description of pattern classes. For an observed feature vector, the Bayesian decision rule is

$$\lambda \;=\; \underset{\kappa}{\operatorname{argmax}}\; p(\Omega_\kappa|\boldsymbol{c}) \;\;=\;\; \underset{\kappa}{\operatorname{argmax}}\; p(\Omega_\kappa)p(\boldsymbol{c}|\Omega_\kappa) \quad , \tag{1}$$

i.e., we decide for that class with highest a posteriori probability. The basic problem for the implementation of statistical classifiers is the definition of adequate a posteriori probabilities. It is a priori not obvious how this statistical concept can be applied to solve 3–D object recognition and pose estimation problems. The required generalization of (1) is guided by the ground rules of Bayesian image analysis approaches stated by Besag (1993), which are commented in the following:

1. *Underlying images, scenes or features have to be characterized by prior probabilities.*

   These statistical measures define the *prior knowledge*; they describe, for instance, the probabilities for the appearance of objects, for the permitted pose parameters or for specific configurations of objects in the scene. The prior knowledge also allows to incorporate prior geometrical information for object recognition. At this point we do not consider the observable features, yet.

2. *Joint probability density functions for observations have to be defined.*

   The statistical behavior of *observable features* has to be defined by a probability density function. This statistical measure describes the probability that a set of features appears, if a special object is present. If the features vary with the object's pose, this density function depends on the position and orientation of objects, too.

3. *Prior probabilities and the joint density functions are combined to find the probability density function.*

   The *combination* of prior probabilities and the feature specific joint density functions results in a probability measure, which can be applied to recognition and pose estimation. This probability measure is called *model density* and describes a traditional form of regularization. The observable features and prior knowledge equally contribute to these model densities, and form the basic mathematical concept for model generation, pose estimation, and classification.

4. *Definition of an inference strategy which allows the efficient computation of a posteriori probabilities for classification.*

   The *evaluation* of a posteriori probabilities is necessary for applying the Bayesian decision rule. Efficient methods are required for the computation of a posteriori probabilities. If hidden Markov models are used, for example, the inference algorithm utilizes the efficient forward–backward algorithm (Rabiner and Juang (1993)).

These guidelines constitute the recipe for the introduction of statistical models for 3–D object recognition purposes.

3 Statistical Modeling of 3–D Objects

The Bayesian framework for 3–D object recognition based on 2–D images has to incorporate the following elements: prior knowledge, rotation and translation of objects, self–occlusion, projection from the model into the image space, and statistical modeling of errors and inaccuracies caused by varying illumination, sensor noise or segmentation errors.

Here, we will not consider single pixels or grid models, but restrict the statistical modeling on segmented images. We assume that the image $[f_{i,j}]$ is transformed into a set of D_{image}–dimensional feature vectors, i.e., the segmentation operator $\mathcal{S}$ defines the mapping

$$\mathcal{S} : [f_{i,j}] \mapsto \boldsymbol{O} \quad , \tag{2}$$

where $\boldsymbol{O} = \{\boldsymbol{o}_k \in \mathbb{R}^{D_{\text{image}}} | 1 \leq k \leq m\}$. Within the segmentation step points, lines, regions or other features can be computed. The number of observed, which are projected to the 2–D image plane is not constant for different images. The cardinality of $\boldsymbol{O}$ depends on the viewing direction, on the applied segmentation algorithm, and on the lighting conditions. Due to the projection, the range information and the assignment between image and model features is lost. The statistical model generation, classification, and localization are limited to these projected feature vectors $\boldsymbol{O}$. In general, model densities of 3–D objects appearing in images embody three principal components: the uncertainty of observed feature vectors, the dependency of features on the object's pose, and the correspondence between image and model features. The statistical description of an

object belonging to class Ω_κ is defined by the density $p(\boldsymbol{O}|\boldsymbol{B}_\kappa, \boldsymbol{R}, \boldsymbol{t})$, and discrete priors $p(\Omega_\kappa)$, $1 \leq \kappa \leq K$, if only single objects appear, or $p(\Omega_{\kappa_1}, \Omega_{\kappa_2}, \ldots, \Omega_{\kappa_q})$ for multiple object scenes. Here, $\boldsymbol{O}$ represents the set of observed feature vectors, and the parameter set $\boldsymbol{B}_\kappa$ contains the model–specific parameters, which model the statistical behavior of features as well as the assignment. The parameters $\boldsymbol{R}$ and $\boldsymbol{t}$, however, symbolize the rotation, translation, and the projection from the model space into the image plane.

The major problem now is the explicit definition of $p(\boldsymbol{O}|\boldsymbol{B}_\kappa, \boldsymbol{R}, \boldsymbol{t})$. Generally, we distinguish between the 3–D model and the 2–D image space. The observable D_{image}–dimensional image features are characterized by $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_m\}$. The corresponding D_{model}–dimensional features in the model space are denoted by $\boldsymbol{C}_\kappa = \{\boldsymbol{c}_{\kappa,1}, \boldsymbol{c}_{\kappa,2}, \ldots, \boldsymbol{c}_{\kappa,n_\kappa}\}$, where in general $n_\kappa \neq m$ due to segmentation errors and occlusion.

Example: *If a 3–D cube is characterized by its corners, $\boldsymbol{C}_\kappa$ includes the 3–D corners. The 2–D image features $\boldsymbol{O}$ are the projected 3–D corners of the model.*

Let us assume the parametric density of the model feature $\boldsymbol{c}_{\kappa,l_k}$ corresponding to $\boldsymbol{o}_k$ is given by $p(\boldsymbol{c}_{\kappa,l_k}|\boldsymbol{a}_{\kappa,l_k})$. A standard density transform results in the density $p(\boldsymbol{o}_k|\boldsymbol{a}_{\kappa,l_k}, \boldsymbol{R}, \boldsymbol{t})$, which characterizes the statistical behavior of the feature $\boldsymbol{o}_k$ in the image plane dependent on the object's pose parameters $\boldsymbol{R}$ and $\boldsymbol{t}$.

Example: *If normally distributed 3–D model features are present, then $\boldsymbol{a}_{\kappa,l_k}$ includes the 3–D mean vector $\boldsymbol{\mu}_{\kappa,l_k}$ and the (3×3) covariance matrix $\boldsymbol{\Sigma}_{\kappa,l_k}$. Let the affine transform $\boldsymbol{o}_k = \boldsymbol{R}\boldsymbol{c}_{\kappa,l_k} + \boldsymbol{t}$ define the mapping from the model into the image space. The image feature $\boldsymbol{o}_k$ is again normally distributed with mean vector $\boldsymbol{R}\boldsymbol{\mu}_{\kappa,l_k} + \boldsymbol{t}$ and covariance matrix $\boldsymbol{R}\boldsymbol{\Sigma}_{\kappa,l_k}\boldsymbol{R}^T$.*

An assignment function ζ_κ defines a discrete mapping, which yields for an observed feature $\boldsymbol{o}_k$ the index $l_k \in \{1, 2, \ldots, n_\kappa\}$ of the corresponding model feature $\boldsymbol{c}_{\kappa,l_k}$, i.e., $\zeta_\kappa(\boldsymbol{o}_k) = l_k$. A set of observed features can thus be associated with the assignment vector $\boldsymbol{\zeta}_\kappa = (\zeta_\kappa(\boldsymbol{o}_1), \zeta_\kappa(\boldsymbol{o}_2), \ldots, \zeta_\kappa(\boldsymbol{o}_m))^T$, which is considered to be a random vector, i.e., the classical matching problem is also modelled statistically. The discrete probability of this random vector is denoted by $p(\boldsymbol{\zeta}_\kappa)$. The probability density function for observing the set of features $\boldsymbol{O}$ thus is,

$$p(\boldsymbol{O}|\boldsymbol{B}_\kappa, \boldsymbol{R}, \boldsymbol{t}) \;=\; \sum_{\boldsymbol{\zeta}_\kappa} p(\boldsymbol{\zeta}_\kappa) \prod_{k=1}^{m} p(\boldsymbol{o}_k|\boldsymbol{a}_{\zeta_\kappa(\boldsymbol{o}_k)}, \boldsymbol{R}, \boldsymbol{t}) \quad , \tag{3}$$

wherein the non observable assignment is eliminated by marginalization, i.e., we sum over all assignments ζ_κ. The evaluation of (3) is computationally bounded by $\mathcal{O}(n_\kappa^m m)$. If pairwise statistically independent assignments are assumed, this complexity reduces to $\mathcal{O}(n_\kappa m)$, and we get a product of density mixtures. Hidden Markov models are derived from (3), if statistically dependent assignments of first order are assumed and the feature transform is omitted. The inference strategy for this case is bounded by $\mathcal{O}(n_\kappa^2 m)$ (Hornegger (1996)).

This flexible formalism of model densities can easily be applied to use multiple views for pose estimation or classification. Assume there are N different views yielding the feature sets ${}^1\boldsymbol{O}, {}^2\boldsymbol{O}, \ldots, {}^N\boldsymbol{O}$. The correct pose parameters are denoted by $\boldsymbol{R}$ and $\boldsymbol{t}$. The images are grabbed by a camera, which is mounted on a

calibrated robot arm. Thus, the approximate extrinsic parameters ${}^{\varrho}\boldsymbol{R}$ and ${}^{\varrho}\boldsymbol{t}$ for each view ${}^{\varrho}\boldsymbol{O}$ are known. These parameters can be expressed in terms of sums using the viewed object's pose $\boldsymbol{R}$ and $\boldsymbol{t}$:

$$ {}^{\varrho}\boldsymbol{R} \;=\; \boldsymbol{R} + \Delta^{\varrho}\boldsymbol{R} \quad \text{and} \quad {}^{\varrho}\boldsymbol{t} \;=\; \boldsymbol{t} + \Delta^{\varrho}\boldsymbol{t} \quad . \tag{4} $$

The density for multiple observations thus is

$$ p({}^{1}\boldsymbol{O}, {}^{1}\boldsymbol{O}, \ldots, {}^{N}\boldsymbol{O} | \boldsymbol{B}_{\kappa}, \boldsymbol{R}, \boldsymbol{t}) \;=\; \prod_{\varrho=1}^{N} p({}^{\varrho}\boldsymbol{O} | \boldsymbol{B}_{\kappa}, \boldsymbol{R} + \Delta^{\varrho}\boldsymbol{R}, \boldsymbol{t} + \Delta^{\varrho}\boldsymbol{t}) \quad , \tag{5} $$

if statistically independent views are presupposed. The use of multiple views will improve the discriminating power of the observed features, because the more data are available for pose estimation and classification, the more reliable results can be expected, even if calibration results will not provide the exact parameters.

4 Statistical Object Recognition

The automatic generation of model densities includes different components: the definition of the structure and the computation of free parameters. The number of model features, the distribution of single features, the mapping from the model into the image space and the dependency of assignments characterize the structure. A practical solution of automatic structure generation is an open research problem (Hornegger (1996)). Nevertheless, there exist algorithms for the estimation of the parameter set $\boldsymbol{B}_{\kappa}$, if the structure of the model density is defined; the computation of $\boldsymbol{B}_{\kappa}$ for each object class Ω_{κ}, $\kappa = 1, 2, \ldots, K$ includes the estimation of the discrete probabilities $p(\boldsymbol{\zeta}_{\kappa})$, which model the assignment function, and $\{\boldsymbol{a}_{\kappa,l} \mid l = 1, \ldots, n_{\kappa}\}$, which characterizes single model features. The available training data consist of features, which are projected model features. The depth information as well as the assignment function are missing. Therefore, the computation of $\boldsymbol{B}_{\kappa}$ corresponds to an incomplete data estimation problem. An established method which can deal with this type of parameter estimation problems is provided by the Expectation Maximization algorithm (Dempster et. al. (1977)). For normally distributed point features, for instance, there exist closed form iteration formulas which allow the estimation of mean vectors from projections. The interested reader will find the complete derivation of several training algorithms for normally distributed point and line features in Hornegger (1996). The probabilistic modeling of objects makes the application of the Bayesian decision rule (1) possible, but some extensions are required. Instead of a single vector a set of features $\boldsymbol{O}$ is given. Furthermore, the pose parameters are part of the probability density functions. The modified Bayesian decision rule, which allows the statistical classification of objects thus is

$$ \lambda \;=\; \underset{\kappa}{\operatorname{argmax}}\; p(\Omega_{\kappa} | \boldsymbol{O}) \;=\; \underset{\kappa}{\operatorname{argmax}}\; p(\Omega_{\kappa}) p(\boldsymbol{O} | \boldsymbol{B}_{\kappa}, \boldsymbol{R}, \boldsymbol{t}) \quad . \tag{6} $$

Since rotation and translation of objects is a priori unknown, $\boldsymbol{R}$ and $\boldsymbol{t}$ are free parameters. A posteriori probabilities $p(\Omega_{\kappa} | \boldsymbol{O})$ cannot be evaluated explicitly.

The pose estimation stage has to compute the best position and orientation before the class decision is possible; the estimation of $\boldsymbol{R}$ and $\boldsymbol{t}$ corresponds to the maximization problem

$$\{\widehat{\boldsymbol{R}}, \widehat{\boldsymbol{t}}\} \quad = \quad \underset{\boldsymbol{R},\boldsymbol{t}}{\operatorname{argmax}}\, p(\boldsymbol{O}|\boldsymbol{B}_\kappa, \boldsymbol{R}, \boldsymbol{t}) \quad . \tag{7}$$

This parameter estimation task is associated with a global optimization problem of a concave multimodal likelihood function. Probabilistic optimization routines are discussed in Hornegger (1996) which allow practically efficient solutions.

5 Experimental Results

The experimental evaluations examine several aspects: we compare standard methods for pose estimation with the introduced statistical approach, show the improvement of pose estimation results using multiple views, and discuss the recognition rates based on a test set including 1600 randomly chosen views of simple polyhedral objects (Figure 1). All experiments run on an HP 9000/735 (99 MHz, 124 MIPS).

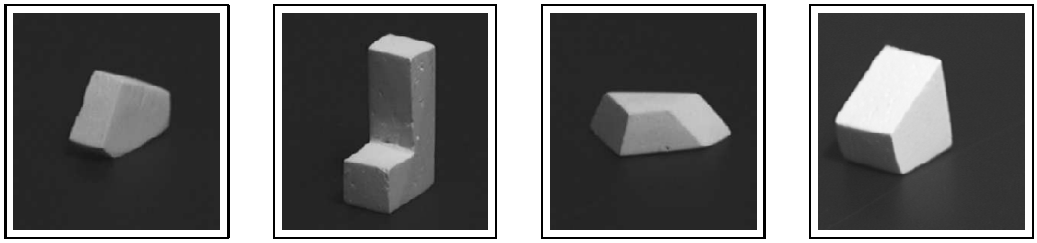

Figure 1: Polyhedral 3–D objects

First a comparison of pose estimation techniques based on the geometrical alignment method of Huttenlocher (1993) and the statistical approach was done using 49 test images. The statistical pose estimation algorithm requires 80s using global optimization, and the alignment method needs 70s in average. The correct pose is computed for 45 images using the statistical approach (see Figure 2). The alignment method failed for 11 images. This experiment shows that the statistical approach can compete with geometrically based methods both with respect to reliability and run time. The computation time for pose estimation is crucially influenced by the global optimization module and its efficiency. The parameter space of continuous model densities is easily partitioned into disjoint subsets which can be considered independent from each other. The use of four processors, for instance, results in a speed–up of 3.5. Table 1 summarizes the speed–up for increasing numbers of processors.

The use of multiple views for pose estimation shows remarkable improvements regarding the correct pose parameters. We run experiments using 400 views and the correct pose parameters increased from 96% to 100%. Existing ambiguities

number of processors	2	3	4	5	6
speed–up	1.7	2.8	3.5	3.9	4.2

Table 1: Parallelization of pose estimation

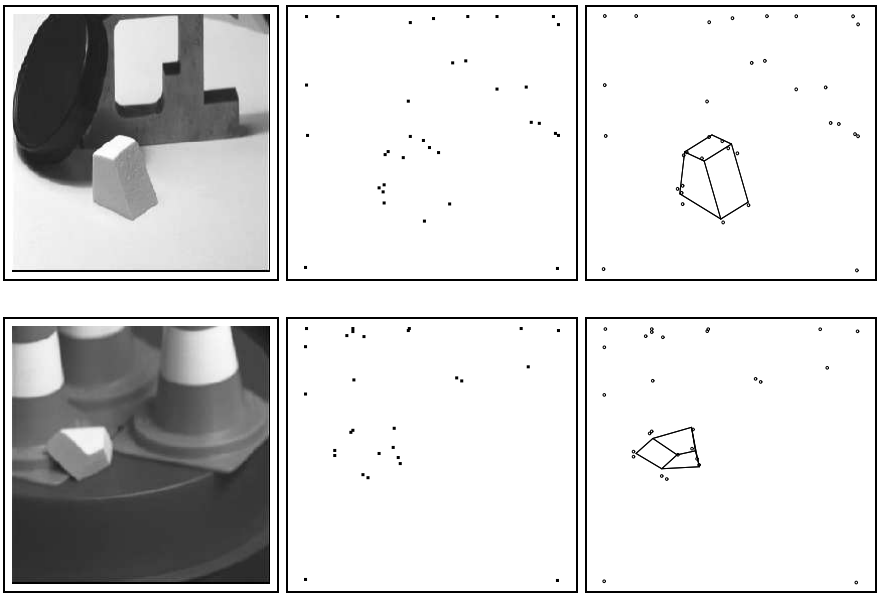

Figure 2: Examples of scenes with heterogeneous background (left: gray–level image, middle: segmentation result, right: estimated pose)

considering a single image are eliminated with a second view, but the average computation time using two views instead of one is three times higher. In average it takes 420s to compute the right position.

The recognition results using 1600 test images of objects shown in Figure 1 are summarized in Table 2. It is distinguished between point and line features.

6 Summary and Conclusions

In this paper, we proposed a framework for Bayesian image analysis. We presented a coherent approach to both modeling single features and the probabilistic characterization of the assignment function between image and model features. The introduced concept of model densities combine assignment, rotation, translation, projection of features, and prior knowledge in a unified manner. The model generation process has to deal with incomplete data estimation problems, whereas the pose computation corresponds to a maximum likelihood estimation. Due to

3–D object	recognition rate [%]		run time per image [sec]	
	points	lines	points	lines
Ω_1	47	44	466	1882
Ω_2	78	82	485	2101
Ω_3	58	36	465	1933
Ω_4	89	76	471	1520
average	68	59	472	1859

Table 2: Run time and recognition rate of 3–D experiments

the statistical nature of the introduced modeling scheme, the implementation of Bayesian classifiers for object recognition is made possible. Experimental results with real data show the practical use of statistical classifiers in computer vision. Indeed, with respect to computer vision applications statistical methods are still in its infancy, but the implemented and evaluated applications show that there is a considerable potential for future development and research.

References

BESAG, J. (1993): Towards Bayesian Image Analysis In: K.V. Mardia and G.K. Kanji (eds.): *Statistics and Images*, volume 1. Carfax Publishing Company, Abingdon, 107–119.

DEMPSTER, A., LAIRD, N. and RUBIN, D. (1977): Maximum Likelihood from Incomplete Data via the EM Algorithm *Journal of the Royal Statistical Society, series B (Methodological), volume 39, number 1, 1–38.*

DUDA, R. and HART, P. (1973): Pattern Classification and Scene Analysis. John Wiley & Sons, New York.

FAUGERAS, O. (1993): Three–Dimensional Computer Vision – A Geometric Viewpoint. MIT Press, Cambridge, Massachusetts.

HORNEGGER, J. (1996): Statistische Modellierung, Klassifikation und Lokalisation von Objekten. Shaker, Aachen.

HUTTENLOCHER, D. (1993): Recognition by Alignment In: A.K. Jain and P.J. Flynn (eds.): *Three–Dimensional Object Recognition Systems*. Elsevier, Amsterdam, 311–324.

RABINER, L. and JUANG, B.–H. (1993): Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, NJ.