A Bayesian Approach to Learn and Classify 3D Objects from Intensity Images

J. Hornegger and H. Niemann

Lehrstuhl für Mustererkennung, Universität Erlangen–Nürnberg,
Martensstr. 3, D–91058 Erlangen, Germany

Abstract

This contribution treats the problem of learning and recognizing 3D objects using 2D views. We present a new Bayesian approach to 3D computer vision based on the Expectation–Maximization–Algorithm, where learning and classification of objects correspond to parameter estimation algorithms. We give a formal description of different learning and recognition stages and conclude the associated statistical optimization problems for each Bayesian decision. The training stage is supposed to be unsupervised in the sense that no explicit feature matching among different views is necessary. Finally, the experimental part of the paper considers the special case, where observable point features are assumed to be normally distributed and the object and its projections are modeled by mixture density functions.

1 Introduction and Motivation

Object recognition systems are often expected to have the capability of learning and recognizing 3D objects from 2D views without the use of previously computed geometric models, like CAD data bases. We present a Bayesian solution to the mentioned learning problem using segmented intensity images. The approach is the first which strictly applies the **E**xpectation–**M**aximization–Algorithm (EM–Algorithm) [1] for that purpose. Both, the learning and the recognition stage of objects are stated as optimization problems, i.e. the maximization of a posteriori probabilities. The algorithms are first described on a high level of abstraction. If the parametric statistical distributions of features are known, the explicit formulas for implementation purposes can be derived, but this causes in general some difficult mathematical problems, i.e. solving definite integrals or the computation of matrix derivatives. The special case of normally distributed model features is discussed and evaluated by experiments. The model features and their projections are therefore represented as a family of mixture density functions (see also [3]). The EM–Algorithm approximates the maximum–likelihood estimates for these incomplete data problems based on the observable projected 2D point features.

2 The Problem

It is a well known result from decision theory that Bayesian classifiers are optimal due to the minimization of the error probability [2]. Hence, the design of a classification system aims to get very close to this theoretical minimum error bound. In the case of 3D object recognition, the classification process includes both the assignment of a subset of observable image features to a model and the computation of the object's pose, described by a rotation matrix and a translation vector. Let $\boldsymbol{C} = \{\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_u\}$ be the set of all available models and let each model $\boldsymbol{C}_\kappa$ be represented by a set of features $\{\boldsymbol{C}_{\kappa,1}, \boldsymbol{C}_{\kappa,2}, \ldots, \boldsymbol{C}_{\kappa,n_\kappa}\}$, where $\kappa \in \{1, 2, \ldots, u\}$. Furthermore, we denote observable image features of the j–th $(1 \leq j \leq J)$ scene with $\boldsymbol{O}_j = \{\boldsymbol{O}_{j,1}, \boldsymbol{O}_{j,2}, \ldots, \boldsymbol{O}_{j,m_j}\}$. Those scene features which do not correspond to any model feature are assigned to the special background feature $\boldsymbol{C}_{\kappa,0}$. Due to the fact that segmentation results show instabilities concerning different views and varying illumination conditions (see Fig.1), in the following both the model features and the observable scene primitives are assumed to be statistical random variables having parameterized density functions. Let the parameters for the model set $\boldsymbol{C}$ be $\boldsymbol{B} = \{\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots \boldsymbol{B}_u\}$. The parameters concerning each model $\boldsymbol{C}_\kappa$ are given by $\boldsymbol{B}_\kappa = \{\boldsymbol{a}_{\kappa,1}, \boldsymbol{a}_{\kappa,2}, \ldots, \boldsymbol{a}_{\kappa,n_\kappa}\}$. The relation between the densities of model and image features is given by a density transform determined by the rotation, translation and the projection properties. Therefore, if a suitable statistical model for each 3D object exists, the process of learning corresponds to the estimation of the parameter set $\boldsymbol{B}$ from the transformed observable random variables. Indeed, the set of training data $\boldsymbol{X}$, whereupon the computation of $\boldsymbol{B}$ is based on, can differ considerably: The sample set may include the matching between features of different views, it may have pose information for each view, or it may contain background features. Thus, we have differnt stages of unsupervised learning processes. The fundamental problem of learning for recognition purposes is the maximization of the probability $p(\boldsymbol{X}|\boldsymbol{B})$ for different types of observations with respect to the parameter set $\boldsymbol{B}$. The recognition of an object modeled by $\boldsymbol{C}_\kappa$ in a given view $\boldsymbol{O}_j$ is also carried out by the maximization of a posteriori probabilities – this time with respect to rotation and translation parameters. In contrast to the learning stage, different classification goals – like multiple object recognition or localisation – cause different optimization problems. The set of observable data is restricted to image features, thereby. The following section will describe an iterative parameter estimation algorithm where the relation between observable image features and the unknown 3D structure of objects is included.

3 Mathematical Framework

We have seen that there may occur several different levels of learning and classification problems concerning the domain of 3D object recognition. The associated statistical optimization problems differ in the data which are either observable or unknown throughout the training or the recognition stage. The learning and classification of 3D objects can thus be considered as a parameter estimation problem from incomplete data. Dempster et al. [1] developed a widely used algorithm for solving those incomplete data estimation problems – the EM–Algorithm. Assume that a measure space $\boldsymbol{M} = \boldsymbol{X} \cup \boldsymbol{Y}$ of complete data is given and a measurable mapping from the complete data $\boldsymbol{M}$ into the incomplete $\boldsymbol{X}$ is defined; let $\boldsymbol{Y}$ denote the *hidden*, i. e. non observable, information. In [1] it is shown that instead of computing a maximum likelihood estimation for $p(\boldsymbol{X}|\boldsymbol{B})$ we can use the conjecture between observable and non–observable data by maximizing the following *Kullback Leibler statistics* regarding the reestimation $\widehat{\boldsymbol{B}}$ of the actual parameters $\boldsymbol{B}$:

$$Q(\boldsymbol{B}, \widehat{\boldsymbol{B}}) = \int \log\, p(\boldsymbol{X}, \boldsymbol{Y} \mid \widehat{\boldsymbol{B}})\, p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{B})\, d\boldsymbol{Y}.$$

Let us summarize the abstract statistical formulation of the 3D object recognition and learning problem: First, we have to model the object's representation in form of a parameterized density function with respect to the parameter set $\boldsymbol{B}$ and the complete data – including observable and non observable components. In a second step, the mapping from complete into the incomplete data domain has to be defined. Finally, the conditioned expectation $Q(\boldsymbol{B}, \widehat{\boldsymbol{B}})$ has to be computed and iteratively maximized until the algorithm converges to a stationary point.

4 Gaussian Object Features

In this section we use the abstract theoretical framework of the previous section and design and implement a 3D object recognition system including the capability of learning objects based on a special model concerning the distributions of features. If a 3D object undergoes a rigid transformation and the locations of the chosen features in the image obey the same rigid mapping, we call those features *attached features*. We suppose that the used features in the 3D space and their corresponding features in the image–plane are simply attached, normally distributed point features. The transformation of model points into image space is constrained to affine functions. For example, the multiplication of a 3D model point with a rotation matrix, addition with the vector, and subsequent scaled orthogonal projection satisfies this restriction. A common problem in 3D computer vision constitutes the phenomenon of occlusion. Therefore, we weight each point feature with its probability of occurring in a 2D projection. The statistical dependency between features is neglected, and it is presumed that the matching among scene and model features is unknown. Each scene feature may correspond to each model primitive with a certain probability. The set of transformed model features of each learning view is considered as a parametric mixture "population" of points. This motivates the description of the complete 3D object by a mixture density function. The probability for observing the set $\boldsymbol{O}_j$ of image features including the object $\boldsymbol{C}_\kappa$ under the rotation $\boldsymbol{R}_j$ and translation $\boldsymbol{t}_j$ by applying the independency assumption is

$$p(\boldsymbol{O}_j|\boldsymbol{R}_j, \boldsymbol{t}_j, \boldsymbol{B}_\kappa) = \prod_{k=1}^{m_j} \sum_{i=0}^{n_\kappa} p(\boldsymbol{C}_{\kappa,i})\, p(\boldsymbol{O}_{j,k}|\boldsymbol{R}_j, \boldsymbol{t}_j, \boldsymbol{a}_{\kappa,i}).$$

During the learning stage it is assumed that the training images include only one object corresponding to a known model $\boldsymbol{C}_\kappa$ with a homogeneous background, and the parameter set $\boldsymbol{B}_\kappa = \{\boldsymbol{a}_{\kappa,1}, \boldsymbol{a}_{\kappa,2}, \ldots, \boldsymbol{a}_{\kappa,n_\kappa}\}$ has to be estimated. The number of model features n_κ for $\boldsymbol{C}_\kappa$ is expected to be given by the user. If the Kullback–Leibler statistics $Q(\boldsymbol{B}_\kappa, \widehat{\boldsymbol{B}}_\kappa)$ is computed for learning 3D objects from J views and if the gradient information is used for the computation of extrema, we get the following training formulas for the Gaussian mixture density: The weight for each model primitive $\boldsymbol{C}_{\kappa,i}$ is iteratively computed by

$$\widehat{p}(\boldsymbol{C}_{\kappa,l}) = \frac{1}{J\,m_j}\sum_{j=1}^{J}\sum_{k=1}^{m_j} p(\boldsymbol{C}_{\kappa,l}\,|\,\boldsymbol{O}_{j,k},\boldsymbol{R}_j,\boldsymbol{t}_j,\boldsymbol{a}_{\kappa,l}),$$

and for the reestimation of the mean vector $\boldsymbol{\mu}_i$ we get with $\boldsymbol{D}_{i,j} = \boldsymbol{R}_j\boldsymbol{K}_i\boldsymbol{R}_j^T$:

$$\widehat{\boldsymbol{\mu}}_i = \left(\sum_{j=1}^{J}\sum_{k=1}^{m_j} p(\boldsymbol{C}_{\kappa,i}|\boldsymbol{O}_{j,k},\boldsymbol{a}_{\kappa,i})\boldsymbol{R}_j^T\boldsymbol{D}_{i,j}{}^{-1}\boldsymbol{R}_j\right)^{-1}$$
$$\sum_{j=1}^{J}\sum_{k=1}^{m_j} p(\boldsymbol{C}_{\kappa,i}|\boldsymbol{O}_{j,k},\boldsymbol{a}_{\kappa,i})\boldsymbol{R}_j^T\boldsymbol{D}_{i,j}{}^{-1}\left(\boldsymbol{O}_{j,k}-\boldsymbol{t}_j\right).$$

Finally, the covariances can be learned by searching successively the zero crossings of

$$\sum_{j=1}^{J}\sum_{k=1}^{m_j} p(\boldsymbol{C}_{\kappa,i}|\boldsymbol{O}_{j,k},\boldsymbol{a}_{\kappa,i})\boldsymbol{R}_j^T\widehat{\boldsymbol{D}}_{i,j}^{-1}(\widehat{\boldsymbol{D}}_{i,j}-\boldsymbol{S})\widehat{\boldsymbol{D}}_{i,j}^{-1}\boldsymbol{R}_j$$

with respect to the components of $\widehat{\boldsymbol{K}}_i$. In this nonlinear function we set $\widehat{\boldsymbol{D}}_{i,j} = \boldsymbol{R}_j\widehat{\boldsymbol{K}}_i\boldsymbol{R}_j^T$ and $\boldsymbol{S} = (\boldsymbol{O}_{j,k}-\boldsymbol{R}_j\boldsymbol{\mu}_i-\boldsymbol{t}_j)(\boldsymbol{O}_{j,k}-\boldsymbol{R}_j\boldsymbol{\mu}_i-\boldsymbol{t}_j)^T$. During the training, the algorithm expects both the observable image features $\boldsymbol{O}_j$ and the affine transformation given by $\boldsymbol{R}_j$ and $\boldsymbol{t}_j$. The computational efficiency of the classification module is more important than the runtime behavior of the training, because the estimation of the parameter set is done once and off–line. In the pose estimation phase, the parameters to be computed are the rotation and translation of the object. The use of the EM–Algorithm provides no closed form solution for the iterative estimation of the pose parameters. Thus numerical optimization techniques are used to maximize the probability function.

5 Experimental Results

The learning process is supported by a robot where a camera is mounted on its hand. The rotation and translation parameters for each view in the learning stage are computed using the position of the calibrated robot's camera. In Fig. 1 three different example training views are shown. For each object we take 50 views to estimate the mean vectors of 3D point features, their covariances, and their weights. The number of components of the mixture density has to be

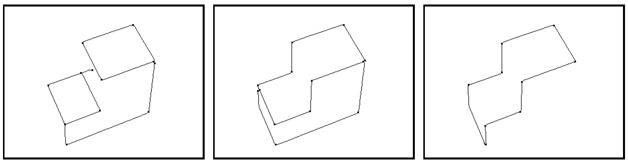

Fig. 1: Segmentation results of different 2D views given to the system. The initialization of means results from the first view, where the depth values for each point are set to zero; all covariance matrices are chosen to be equal at the beginning of the training module and all features are assumed to be uniformly weighted. Dempster et al. [1] propose that the convergence of the EM–Algorithm is very slow. In our experiments 10 iterations were necessary in the worst case for the convergence of the EM learning procedure. The computation of the global maximum of the density function with respect to pose parameters requires actually 5–10 minutes on a HP 735 Workstation (124 MIPS) using an adaptive random search technique, wherein explicit feature matching is avoided.

6 Conclusions

The experiments show that the developed statistical approach to the recognition problem of 3D objects provides promising results for the selected examples. In future, we have to find more sophisticated statistical models for 3D objects and their related projections. The effect of occlusion is not satisfiably modeled in the actual mixture density assumption. Further research should also be concentrated on robust parameter estimation techniques using a limited set of sample data. The initialization of the parameter at the beginning of the iterative optimization will also be an essential factor for the improvement of the proposed statistical approach. The suggested technique using the EM–Algorithm seems to allow a substantial contribution for building robust 3D object recognition systems with the capability of automatic learning from examples. This optimism is moreover motivated by the success in the field of speech recognition using the concept of Hidden Markov Models which is a special case application of the EM–Algorithm [2].

7 References

1. A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.
2. H. Niemann. *Pattern Analysis and Understanding.* Springer, Heidelberg, 1990.
3. W. M. Wells III. Statistical Object Recognition. PhD thesis, MIT, Department of Electrical Engineering and Computer Science, Massachusetts, February 1993.