Bayesian Vision: From Intensity Marginals to Mutual Information and Entropic Object Recognition

Abstract

In this paper, we introduce a new and general framework for active statistical object recognition. Model generation, classification, localization, as well as viewpoint planning are considered in a unified manner using either gray–level images or derived features. Instead of introducing geometrically motivated object models, the representation of objects is based on density functions. These include both structural and statistical information on objects and their appearance in the image plane. If no pose–invariant features are used, statistical object models will include rotation and translation parameters as well as characteristic properties of projections from the model space into the image plane. Spatial distributions of single intensity values or image features are estimated using a set of sample views and the expectation maximization algorithm. This method can deal with latent training data, i.e., missing depth and unknown assignments. The probabilistic framework permits the application of powerful mathematical tools that reduce the complexity of recognition, localization, and viewpoint planning algorithms: mutual information, density transformations and marginalizations. Impressive examples show the advantages of the introduced statistical concepts applied to object recognition problems.

Keywords: statistical image modeling, statistical object recognition, pose estimation, intensity marginal, mixture density, mutual information

1 Introduction

The ultimate goal of object recognition systems is the efficient transformation of image data into symbolic data which describe the present objects and their pose in such a way that the task of the system can be solved in an accurate manner. The application of *statistical* and *active* methods to solve this computer vision task is of increasing interest, and seem to overcome some problems of purely geometrically based techniques with sensor noise, instabilities of images, and the selection of most discriminating views [7, 11, 19, 28]. There exist several reasons which make the use of probabilistic techniques appropriate and suggest the development of a statistical framework for computer vision:

- the exceptional success of statistical methods in other fields of pattern recognition like speech recognition [17],
- instabilities of image features due to varying illumination, quantization errors, or sensor noise [28],
- optimality of the Bayesian decision rule with respect to misclassification in the presence of a 0–1 cost function [4], and
- theoretical results concerning information theory, parametric and non–parametric estimation theory which might be advantageous for model generation algorithms and pose estimation [6, 22].

Many object recognition methods are based on single views and associated features like 2–D points or lines. The instabilities and inaccuracies of features as well as the change of viewing directions are often not sufficiently considered within the chosen mathematical frameworks [10]. Even statistical identification and pose estimation algorithms make use of automatically computed segmentation results [7, 18, 24, 20, 28], but do neither influence the selection of viewpoints

and the number of views required for reliable recognition, nor do they select the segmentation algorithms and features from the set of available functions and values. The statistical modeling of observable image features, of the assignment function or of relational dependencies between observed primitives is done by joint probability density functions. The disadvantages of segmentation–based approaches using single views are manyfold: segmentation algorithms lead to the reduction of data which usually results in features of lower information and discriminating power. The more data are available, the higher is the expected discriminating power. This is also valid for the sensor data. If the chosen camera position is not fixed and the algorithm can select a viewing direction which shows the highest discriminating power for the given object, we expect more reliable recognition results. For traditional approaches to object recognition based on segmentation, it is true that — even if the segmentation results are modelled statistically — illumination and other properties are not part of the probabilistic model, explicitly.

Instead of using segmentation results and one single, randomly chosen view for recognition, it seems natural to use intensity values or responses of selected filtering operations directly. The viewing direction should be selected such that ambiguities are resolved and a reliable identification is possible. For that reason and due to the motivations mentioned within the introduction, this paper introduces spatially dependent parametric density functions for image modeling and illustrates, how these probabilistic models can be used for efficient pose estimation and classification. The statistical framework allows also the judgment of views in an information–theoretical manner. For a given viewing direction, the amount of mutual information between the model and the observation can be measured using the available probability density functions, the pose parameters, and the gray–level image. Also the usage of multiple views for recognition purposes is possible.

The basic problems treated in this paper are summarized as follows:

- statistical modeling of objects and their appearance in the image plane,
- computation of most discriminating views with respect to the given model data base,
- dynamic determination of the most rational recognition algorithm, and
- use of marginals to reduce the computational complexity of pose estimation and viewpoint selection.

The remainder of this paper is organized as follows: next section gives a discussion of related work and a brief summary of statistical methods for object recognition and state–of–the–art for viewpoint planning. The third section introduces the theoretical framework for statistical image modeling and shows selected examples. These include mixture models and the embedding of pose parameters into model densities. We also discuss the mathematical framework for model generation algorithms (Section 4), pose estimation methods (Section 5), and the classification problem (Section 6). Section 7 deals with the complexity of statistical object recognition and pose estimation, and shows that projections and associated marginals lead to simplifications which induce more efficient recognition algorithms. Marginalization is considered as a new and powerful method which decreases the complexity for pose estimation and classification. The viewpoint planning problem using information theory is part of Section 8. We show how views of highest discriminating power can be selected automatically using mutual information. The paper concludes with the experimental evaluation of introduced concepts and a brief summary, which also gives some hints to future research problems.

2 Related Work

Most object recognition systems based on gray–levels apply averaging techniques. Invariant features for the identification of 2–D objects are computed [14] or some histogram type representations are chosen [19]. Often the pose estimation problem is of minor interest, and the representation of unknown objects in cluttered background is not part of the modeling formalism. A first gray–level based approach to estimate pose parameters of 3–D objects using mutual information is introduced in [26]. This algorithm requires a 3–D model, and applies methods of information theory and non–parametric densities for image registration purposes. An appearance based vision system, which applies Karhunen–Loève transform to image vectors, is discussed in [15]. Each view is associated with a vector, the complete object is represented by a manifold induced by feature vectors of different views. A recommended comparison of appearance based approaches with other representations can be found in [16]. The discussion there shows that appearance based methods, which avoid segmentation, lead to more robust recognition modules than methods which are restricted to geometric primitives, like point or line features. This result has also motivated the approach proposed in this paper.

The selection of viewpoints depends on two facts [25, 23]: on the one hand the view can be selected which is the most probable and stable view of an object, and on the other hand the viewpoint shows the 2–D projection which allows the best distinction of considered objects with respect to the model database. The mathematical fundamentals for likelihood of views and view stability of single objects are introduced and experimentally verified in [27]. This method allows the definition of prior probabilities for different viewing directions, and thus reduces the search space for pose parameters. These measures only depend on the object and do not consider other objects of the model database. In contrast to established techniques for viewpoint

planning [23], the consideration of pattern recognition problems using information theory, as discussed in [6], can be extended to compute those features with highest discriminating power. The maximization of mutual information was first proposed by Schiele and Crowley [19] to viewpoint planning and robot vision applications, who report remarkable experimental tests on this method using multiple perceptive fields. Here we extend these ideas to a more general setting.

3 Statistical Modeling of Intensity Images

Bayesian image analysis and viewpoint planning based on information theory requires statistical descriptions of objects and their appearance in the image plane. For that reason, we are looking for a general mathematical framework which allows the definition of *model densities*. This can be, for instance, achieved by the registration of probabilistic properties of gray–levels or by the statistics of observable geometric primitives like 2–D points or regions [7, 12]. The required probabilistic models here should allow both identification and localization of objects. Thus, a mathematical formalism is needed which yields a theoretical framework for the statistical description of the spatial behavior of intensity values or image features dependent on the objects' pose. Position and orientation are related to a pre–defined reference coordinate system. First, the discussion concentrates on the statistical modeling of objects using intensity images. Second, specializations will result in model densities for more abstract features, and show the power and generality of the introduced formalism. The given examples stress the generality of the chosen statistical modeling scheme.

3.1 Mixture Modeling of Spatial Distributions

An intensity image $\boldsymbol{f} = [f_{i,j}]_{1 \leq i \leq N, 1 \leq j \leq M}$ is represented as a matrix of discrete values, which are typically gray–level or color images. For simplicity, the following discussion is restricted to scalars, such as gray–levels. Within a statistical setting, the complete observable image is considered as a random field. There are several possibilities to characterize the statistical behavior of intensity values. Histograms, for example, are suitable for the description of gray–levels and their relative frequencies. They are successfully applied to object identification [2] and localization (e.g., histogram backprojection, [21]). But, in general, histograms do neither reflect the spatial distribution of intensity values in the image plane, nor the dependency of gray–levels on the object's pose parameters. For that reason, Markov random fields [11] or hidden Markov mesh fields [3] are widely used for image modeling. These statistical representations include the spatial distribution of gray–levels as well as spatial dependencies of considered random variables. The geometrical 3–D structure of objects and the relation to 2–D projections, however, is not explicitly represented, in contrast to structural descriptions, such as in [5].

The proposed statistical modeling scheme considers spatial distributions of single intensity values in the image plane dependent on the gray–levels, and models these by bivariate probability density functions (p.d.f.). A parametric representation of p.d.f.'s can be obtained by mixtures of Gaussians. Gaussians are adequate, because a well–known theoretical result states that linear combinations of Gaussians allow the approximation of arbitrary p.d.f.'s up to a certain error bound [29].

For the mathematical formalization, let us assume that intensity values are discrete and can have r different values, i.e., $g_1, g_2, \ldots, g_r$. Usually a quantization of eight bits is assumed, and thus g_l is an element of the set $\{0, 1, \ldots, 255\}$. For each gray–level g_l $(1 \leq l \leq r)$ we

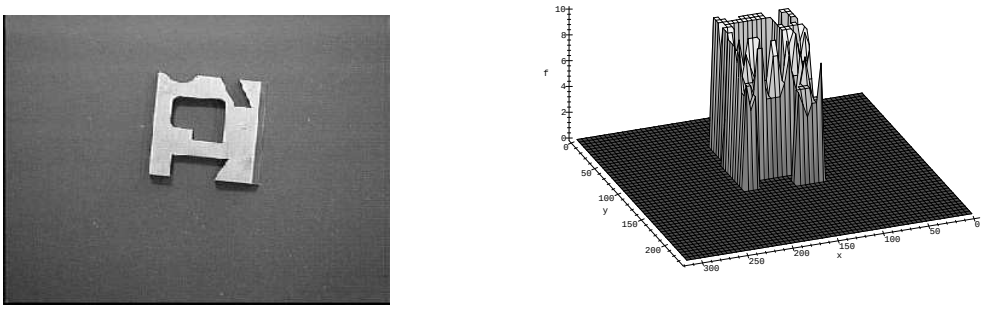

Figure 1: Spatial appearance of different gray–levels. The left image shows a gray–level image, the right plot characterizes the spatial distribution of gray–level 100 using 100 example views.

consider the parametric spatial density function in the image plane $p((i,j)|\boldsymbol{a}_{g_l})$, where $(i,j) \in \{1,2,\ldots,N\} \times \{1,2,\ldots,M\}$ is a 2–D image point; the symbol $\boldsymbol{a}_{g_l}$ denotes the parameters of the density function associated with intensity value g_l which is independent of the lattice point (i,j). The observable image is decomposed into r images, one image for each intensity value g_l, $1 \leq l \leq r$. At this point, the modeling of these images is done seperately.

Example: *Figure 1 shows a gray–level image and the density for a single intensity value. The spatial distribution of intensity value g_l can be approximated by a mixture of Gaussians. In this case, the parameters $\boldsymbol{a}_{g_l}$ summarize the characteristic parameters of the mixture, i.e., the discrete probability for each mixture component, the mean vectors, and the covariance matrices. If r intensity values are present, we get r images, one for each gray–level, and thus r mixtures of Gaussians.* □

If we assume pairwise statistically independent intensity values, the statistical description of the complete image $\boldsymbol{f}$ defined on the 2–D grid $\boldsymbol{X} = [i,j]_{1 \leq i \leq N, 1 \leq j \leq M}$ is the conditional

probability density function defined by the product

$$p(\boldsymbol{X}\,|\,\boldsymbol{f};\{\boldsymbol{a}_{g_1},\boldsymbol{a}_{g_2},\ldots\boldsymbol{a}_{g_n}\}) \;=\; \prod_{i=1}^{N}\prod_{j=1}^{M} p((i,j)^T|\boldsymbol{a}_{f_{i,j}}) \quad . \tag{1}$$

This density allows the computation of a statistical measure for an observed image, presupposed the parameters of the involved mixtures are known. Intensity values corresponding to objects depend on the pose parameters of objects in the world coordinate system as well as on illumination conditions. Thus, above probability density function (1) has to be extended with respect to these degrees of freedom. This is done by the introduction of density transforms.

3.2 Mixtures with Integrated Feature Transform

Rotation and translation of objects induces a transform of observed random variables in the image plane. Let us assume, a random vector $\boldsymbol{x}$ with p.d.f. $p_{\boldsymbol{x}}(\boldsymbol{x})$ is mapped by $\boldsymbol{T}$ to the random variable $\boldsymbol{y}$. If the transform $\boldsymbol{T}$ is bijective, i.e., the inverse mapping $\boldsymbol{T}^{-1}$ exists, the p.d.f. $p_{\boldsymbol{y}}(\boldsymbol{y})$ of random variable $\boldsymbol{y}$ is [1]:

$$p_{\boldsymbol{y}}(\boldsymbol{y}) \;=\; |\det(J_{\boldsymbol{T}^{-1}}(\boldsymbol{y}))|\, p_{\boldsymbol{x}}(\boldsymbol{T}^{-1}(\boldsymbol{y})) \quad , \tag{2}$$

wherein $J_{\boldsymbol{T}^{-1}}(\boldsymbol{y})$ denotes the Jacobian of $\boldsymbol{T}^{-1}$ at $\boldsymbol{y}$.

Example: *Let $\boldsymbol{x}$ be normally distributed with mean vector $\boldsymbol{\mu}_{\boldsymbol{x}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. An affine bijective mapping is given by the matrix $\boldsymbol{R}$ and the vector $\boldsymbol{t}$, and we define $\boldsymbol{y} = \boldsymbol{R}\boldsymbol{x} + \boldsymbol{t}$. The application of formula (2) shows that the resulting random variable is again normally distributed. The mean vector is $\boldsymbol{\mu}_{\boldsymbol{y}} = \boldsymbol{R}\,\boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{t}$ and for the covariance matrix we get $\boldsymbol{\Sigma}_{\boldsymbol{y}} = \boldsymbol{R}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{R}$.* □

Using this result, we can extend p.d.f. (1) with respect to affine transforms, which include rotations and translations as special cases. The general p.d.f. including additional degrees of freedom thus is

$$p(\boldsymbol{X}|\boldsymbol{f}, \{\boldsymbol{a}_{g_1}, \boldsymbol{a}_{g_2}, \ldots \boldsymbol{a}_{g_n}\}, \boldsymbol{R}, \boldsymbol{t}) \;=\; \prod_{i=1}^{N} \prod_{j=1}^{M} p((i,j)^T | \boldsymbol{a}_{f_{i,j}}, \boldsymbol{R}, \boldsymbol{t}) \quad . \tag{3}$$

Example: *Let us assume that the spatial probability density $p((i,j)|\boldsymbol{a}_{g_l})$ of single intensity values g_l is a mixture of Gaussians, i.e.,*

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{a}_{g_l}) &= \sum_{k=1}^{m_l} p_{g_l,k}\; \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{g_l,k}, \boldsymbol{\Sigma}_{g_l,k}) \\ &= \sum_{k=1}^{m_l} \frac{p_{g_l,k} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{g_l,k})^T \boldsymbol{\Sigma}_{g_l,k}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_{g_l,k})\right)}{\sqrt{\det\left(2\pi \boldsymbol{\Sigma}_{g_l,k}\right)}} \quad , \end{aligned} \tag{4}$$

wherein $\boldsymbol{x}^T = (i,j)$, m_l the number of mixture components corresponding to gray–level g_l, and $\sum_{k=1}^{m_l} p_{g_l,k} = 1$. If we extend these mixtures with respect to affine transforms, we get:

$$p(\boldsymbol{X}|\boldsymbol{f}, \{\boldsymbol{a}_{g_1}, \boldsymbol{a}_{g_2}, \ldots \boldsymbol{a}_{g_n}\}, \boldsymbol{R}, \boldsymbol{t}) \;=\; \prod_{i=1}^{N} \prod_{j=1}^{M} \left(\sum_{k=1}^{m_l} p_{f_{i,j},k}\; \mathcal{N}(\begin{pmatrix} i \\ j \end{pmatrix} | \boldsymbol{R}\boldsymbol{\mu}_{f_{i,j},k} + \boldsymbol{t}, \boldsymbol{R}^T \boldsymbol{\Sigma}_{f_{i,j},k} \boldsymbol{R}) \right) \tag{5}$$

This p.d.f. allows the computation of a probability measure for an image corresponding to an object with pose parameters, here defined by $\boldsymbol{R}$ and $\boldsymbol{t}$. □

Instead of affine mappings, there are several other transformations which might be considered. The rotation, translation, and the subsequent projection of 3–D objects into the image plane, the underlying illumination model, the transformations required for resolution hierarchies or the application of some filtering operations, like corner detection, induce transformations which have to be part of the density functions. In general, most transformations, like

projections, are not bijective mappings, i.e., there exists no inverse. For these transformations the standard density transform given by (2) cannot be applied directly. To overcome this problem, the considered transform has to be extended to a bijective transform. This procedure adds some auxiliary random variables, which can be eliminated in a second step by marginalizations after the density transform [1].

Example: *Let us consider the spatial distribution of point features instead of intensity values. In terms of the above suggested statistical modeling scheme, the p.d.f. for 2–D points requires only a product of a single mixture of densities, because instead of r intensity values, we have only one type of points and the associated positions in the image plane. Figure 2 shows an example for a gray–level image and the result of a standard corner detection algorithm. Compared with intensity values, the appearance of point features is quite sparse in the image plane. Therefore, the discriminating power of point features is expected to be much lower than gray–levels. We assume that the observable point features in the image plane* $\boldsymbol{o}_1$, $\boldsymbol{o}_2, \ldots,$ $\boldsymbol{o}_q$ *are transformed and projected corners of the original 3–D object. Let the involved mapping be defined by the affine transform*

$$\boldsymbol{o}_i \;=\; \boldsymbol{R}\boldsymbol{c} + \boldsymbol{t} \quad , \tag{6}$$

wherein $\boldsymbol{o}_i \in \mathbb{R}^2$, *the corresponding model point* $\boldsymbol{c} \in \mathbb{R}^3$, $\boldsymbol{R} \in \mathbb{R}^{2\times 3}$ *and* $\boldsymbol{t} \in \mathbb{R}^2$. *The statistical distribution of the observable 2–D features can be characterized by a mixture of Gaussians [7, 28]. If we assume that this holds also for the 3–D model points, we get the following p.d.f. for the 2–D observations:*

$$p(\boldsymbol{O}|\boldsymbol{a}, \boldsymbol{R}, \boldsymbol{t}) \;=\; \prod_{i=1}^{q} \sum_{k=1}^{m} p_k \; \mathcal{N}(\boldsymbol{o}_i | \boldsymbol{R}\boldsymbol{\mu}_k + \boldsymbol{t}, \boldsymbol{R}^T \boldsymbol{\Sigma}_k \boldsymbol{R}) \quad , \tag{7}$$

where $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_q\}$, $\boldsymbol{\mu}_k \in \mathbb{R}^3$, *and* $\boldsymbol{\Sigma}_k \in \mathbb{R}^{3\times 3}$. *This p.d.f. shows that the statistical models used in [28] and [7] are specializations of (1). The correspondence problem is not present, since it is eliminated by summation over all mixture components.* □

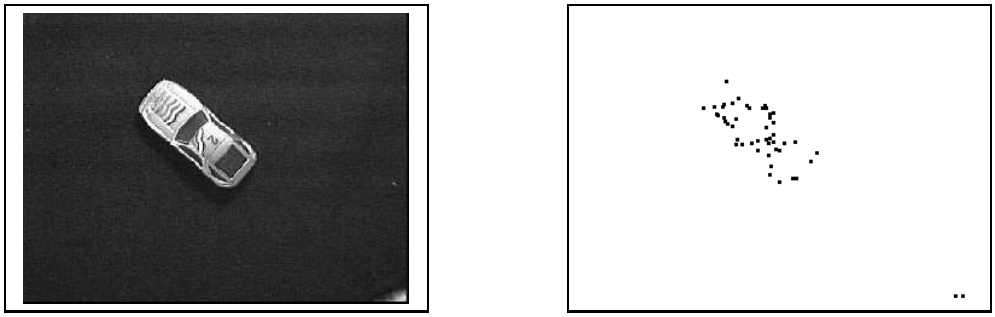

Figure 2: Gray–level image and computed point features

The introduction of p.d.f. with integrated transforms for image modeling raise up several types of problems, which we will consider next:

- How can we estimate the density parameters using sample views?
- How can we estimate the object's pose parameters?
- Which decision rule will be applied for classification?
- Which methods can be applied to increase the efficiency of pose estimation and classification?
- Which methods can be used to viewpoint planning in the presence of statistical models?

We start the discussion of these questions with the model generation stage for statistical object models.

4 Model Generation

The model generation based on model densities corresponds to a parameter estimation problem. For a set of observed training images and the corresponding pose parameters, the model

parameters have to be computed such that the p.d.f. fits the observation with respect to an optimality criterion. If we apply the maximum likelihood estimation for that purpose, the unknown parameters maximize the likelihood function for a given observation.

Example: *Let us assume we approximate the density function for each intensity value g_l using a mixture of Gaussians. The unknown parameters of this model density are (c.f. (5))*

1. *the number m_l of mixture components,*
2. *the discrete probabilities $p_{g_l,k}$, $(1 \leq k \leq m_l)$,*
3. *the m_l 2–D mean vectors, and*
4. *the (2×2) covariance matrices.*

If we have V training views $\boldsymbol{f}_1$, $\boldsymbol{f}_2, \ldots,$ $\boldsymbol{f}_V$ and the corresponding transformation parameters $\boldsymbol{R}_1, \boldsymbol{t}_1, \ldots, \boldsymbol{R}_V, \boldsymbol{t}_V$, the maximum likelihood estimation corresponds to the optimization problem

$$\underset{m_l, \left\{\boldsymbol{\mu}_{g_l,k}, \boldsymbol{\Sigma}_{g_l,k}; 1 \leq k \leq m_l\right\}}{\operatorname{argmax}} \prod_{v=1}^{V} \prod_{i=1}^{N} \prod_{j=1}^{M} \left(\sum_{k=1}^{m_l} p_{f_{v,i,j},k} \, \mathcal{N}(\begin{pmatrix} i \\ j \end{pmatrix} | \boldsymbol{R}_v \boldsymbol{\mu}_{f_{v,i,j},k} + \boldsymbol{t}_v, \boldsymbol{R}_v^T \boldsymbol{\Sigma}_{f_{v,i,j},k} \boldsymbol{R}_v) \right). \tag{8}$$

□

This example shows that we have to deal with two different types of optimization: dynamic and static. The number of mixture components m_l defines the static structure of the density and sets the dimension of the search space for other parameters: the higher we choose m_l, the more mean vectors, covariance matrices, and weights have to be estimated. The computation of the optimal number of mixtures m_l is a dynamic optimization problem, as it is well known from control theory. If m_l is fixed, the static optimization is restricted to the estimation of mixture parameters, like means or covariances. Due to the fact that it is not known in advance which image point corresponds to which component of the chosen mixture, the parameter computation is associated with an incomplete data estimation problem [22].

4.1 Vector Quantization

The initialization of mixtures, which includes the estimation of mixture components and other parameters, is also one of the central problems in speech processing [8]. Usually, speech sample data are pre–processed by a vector quantization step. This reduces the set of all sample vectors to a set of reference vectors of much lower cardinality, which is the so–called *code–book*. Each vector of the sample data is mapped on a vector of the code–book in an uniqe manner.

This method is also applied herer for image processing. For our application, we use the well–known LBG vector quantization algorithm combined with mean square errors to compute the number of mixture components as well as initial estimates of mean vectors [8]. This method, however, is restricted to those application where the dimensions of code–book vectors and sample data are equal. Vector quantization methods which work with projected observations are still open research problems. The estimation, for instance, of a code–book including 3–D vectors using 2–D projected sample data is not possible using these methods.

4.2 Parameter Estimation using Incomplete Data

The estimation of parameters has to be done using training data with missing information. If, for instance, only 2–D projections of 3–D objects can be observed, the depth information is not part of the training set. An established algorithm, which can deal with incomplete training data, is the expectation maximization algorithm (EM algorithm) [22], which is a local optimization method. The EM algorithm is an iterative version of the maximum likelihood estimation. The advantage of the EM algorithm is that for most applications dealing with mixture densities, the search space can be decomposed into lower dimensional and independent sub–spaces. Therefore, the parameter estimation problem is divided into simpler and independent optimization tasks.

Example: *Let us assume we observe sets of 2–D points which are projections of normally distributed 3–D points, and we have to estimate the 3–D mean vectors using the 2–D image points. Due to the fact that we do neither know the depth values, nor the assignment of the 2–D points to the corresponding 3–D points, we have to deal with an incomplete data estimation problem. The application of the EM algorithm results in iterative training formulas which can estimate the mean vectors from projections, without knowing the correspondence between projected points and the missing 3–D information [7]. For a detailed discussion on applications of the EM algorithm with respect to mixtures of densities we recommend [13].* □

5 Pose Estimation

Model densities show two different types of parameters: model parameters, which specify the spatial distribution, and pose parameters, which characterize the position and orientation within the world coordinate system. The model parameters are estimated during the training stage. If the p.d.f. has to be evaluated for a given image $\boldsymbol{f}$, also the knowledge of pose parameters is necessary. Within the statistical framework, this corresponds to the maximum likelihood estimation problem:

$$\{\widehat{\boldsymbol{R}}, \widehat{\boldsymbol{t}}\} \quad = \quad \underset{\boldsymbol{R},\boldsymbol{t}}{\operatorname{argmax}}\, p(\boldsymbol{X}\,|\,\boldsymbol{f}, \{\boldsymbol{a}_{g_1}, \boldsymbol{a}_{g_2}, \ldots \boldsymbol{a}_{g_n}\}, \boldsymbol{R}, \boldsymbol{t}) \quad . \tag{9}$$

Example: *The localization of 2–D objects requires the estimation of the rotation angle in the image plane and the 2–D translation vector. The pose estimation problem is thus related to a global optimization within a 3–D search space.* □

The computation of pose parameters, however, is a computationally hard problem and

requires efficient implementations. We will outline some methods to solve this task in Section 7.

6 Classification

If there are K object classes, which can appear in the image, the classification problem is to find a discrete function which maps a given image $\boldsymbol{f}$ to the correct object class Ω_κ, $\kappa = 1, 2, \ldots, K$. The identification of objects within a statistical framework is based on the Bayesian decision rule. This decision method requires the computation of posterior probabilities. Let $p(\Omega_\lambda|\boldsymbol{f})$ be the a posteriori probability for class Ω_λ, if the image $\boldsymbol{f}$ is given. We decide for that class Ω_κ, which maximizes the a posteriori probability, i.e.,

$$\Omega_\kappa \quad = \quad \underset{\Omega_\lambda}{\operatorname{argmax}}\, p(\Omega_\lambda|\boldsymbol{f}) \quad . \tag{10}$$

This decision rule guarantees the optimality of the implemented classifier with respect to missclassifications if the statistical models are appropriate. The a posteriori probabilities are easily computed, if the pose parameters are known (c.f. Section 5).

7 Methods to Increase Efficiency

Above discussion shows that in the presence of non–invariant features, it is necessary to estimate the pose parameters for recognition purposes. For that reason, the efficiency of recognition algorithms crucially depends on the efficiency of pose estimation. The run–time of the parameter estimation module, which is based on global optimization algorithms, is essentially influenced by

- the dimension of the search space,

- the number of function evaluations, and
- the time required for density evaluations.

The following subsections present some possibilities to reduce the computational effort for pose estimation with respect to these three items.

7.1 Parallelization

Pose estimation within the chosen statistical framework corresponds to a global optimization problem. If, for instance, orthographic projection is assumed, the search space has five dimensions: three rotation angles and two components of the translation vector. The search space, in general, can be partitioned into disjoint subsets. This also induces a decomposition of the search problem into independent sub–tasks, and allows the distribution of the global optimization problem to several processors. For each element of the partition we global maximum. A comparison of global maxima of subspaces results in the global maximum we are looking for. The speed–up depends on the number of used processors and is expected to be linear. Indeed, parallelization is the most obvious acceleration of pose estimation, but there are also some more sophisticated methods, which work on image data and take advantage of some mathematical properties of present objective functions, i.e., model densities.

7.2 Resolution Hierarchies

Algorithms in image processing and computer vision apply resolution hierarchies to reduce the complexity of the considered problem. Lower resolutions reduce the number of image points and therefore the time required for density evaluations. The usage of resolution hierarchies is also advantageous within the context of the chosen statistical framework. The switch between several resolution levels has to be done using statistical models. The involved p.d.f. can be

computed using standard density transforms (c.f. (2)).

Example: *The easiest way to define resolution hierarchies (which is from a theoretical point of view not correct, but works fine on images captures with CCD devices) is to use average gray–levels of a well–defined neighborhood of a given pixel of the image grid. The computation of means of gray–levels induces a density transform on the original p.d.f. (1). Since densities which correspond to sums of random variables can be computed by subsequent convolutions of the summands' densities [1], the p.d.f. of lower resolutions are known. For example, if normally distributed random variables are considered, the convolution operations result in Gaussians, too [1].* □

Instead of using resolution hierarchies, we can also choose only each n–th point of the image grid for density evaluations. A probabilistic method will use a random process to select the image points, which might be considered for density evaluations.

7.3 Quantization

In the previous section, we suggested to reduce the spatial resolution of the image to speed–up the evaluation time of involved density functions. Another reduction can be achieved by the quantization of considered intensity values. Threshold operations, like binarization or histogram linearization, allow the reduction of gray–levels appearing in the image. Also for quantization purposes, a straightforward density transform can be used to compute the p.d.f. of reduced gray–levels.

Example: *A multi–thresholding operation makes it possible to map a gray–level image with 255 gray–levels to an image with, for instance, five gray levels. Figure 3 shows an example of a gray–level image with reduced intensity values. The spatial distribution of five gray–levels can be used as p.d.f. for recognition and reduces the computational complexity of (1), obviously.*

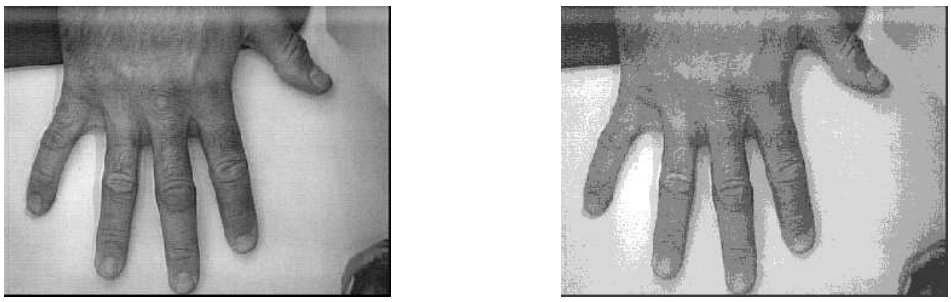

Figure 3: Gray–level image with 256 intensity values (left) and five intensity values (right)

The introduction of features like points or lines is another example for (much more difficult) quantizations, which simplify the original model density (1). □

7.4 Marginals

Up to now, we have partitioned the search space, have reduced the number of function evaluations as well as the complexity of model densities. Methods which reduce the dimension of the search space were not discussed and are not obvious. Figure 4 illustrates that projections can eliminate free parameters. One–dimensional projections of point features, for example, are invariant with respect to translations along the y–axis of the image coordinate systems and regarding rotations around the x–axis. The search space is thus reduced, if one–dimensional point features are considered instead of 2–D points. The consequence for practical implementations is that for pose estimation, we first compute hypotheses in the lower dimensional sub–space and refine these parameters in the higher–dimensional search space. Within the statistical models, projection on the y–axis corresponds to the computation of the marginal density of the original p.d.f. Marginalizations, therefore, simplify the global optimization problems, which are related to pose estimation.

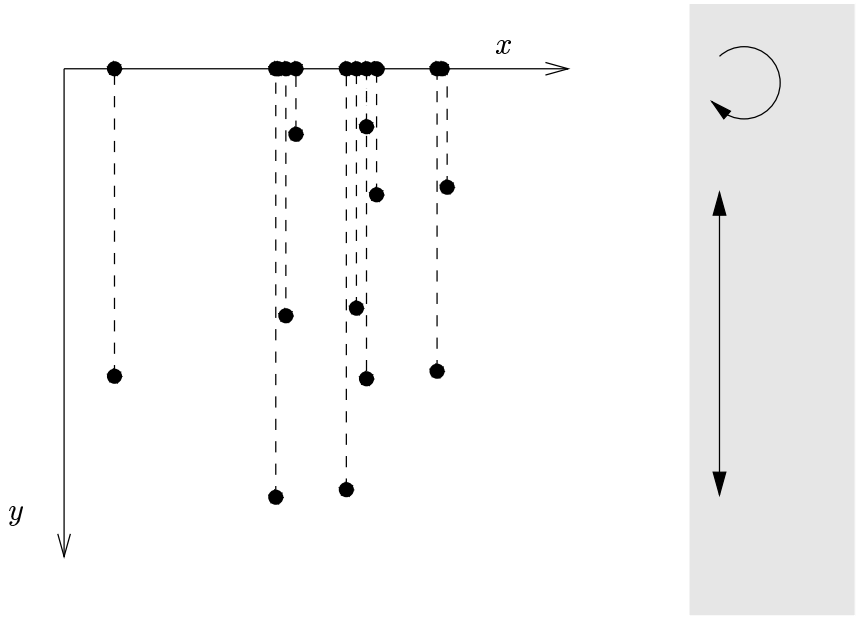

Figure 4: Projections reduce the number of free parameters

Example: *If the spatial distribution of intensity values is modeled by (1), intensity marginals can be computed by integrating out the x– or the y–components. The resulting* semi–invariants *can be applied to solve several problems. Applications of intensity marginals can be found, for instance, in document analysis systems. Therein, marginals are used to detect lines of written text.* □

The Bayesian approach to computer vision presented so far can be used to learn, to recognize, and to localize objects. Nevertheless, the mathematical framework shows some remarkable disadvantages: The projection of features does not necessarily keep the global maximum, and, of course, decreases the discriminating power of features. Furthermore, it might happen that different 3–D objects share the same features in projections. Figure 5, for example, shows two different 3–D objects with a common 2–D view. The computation of pose parameters corresponds to a parameter estimation problem. Due to the fact that consistent estimators show a convergence in probability for increasing sample data, the use of multiple views should

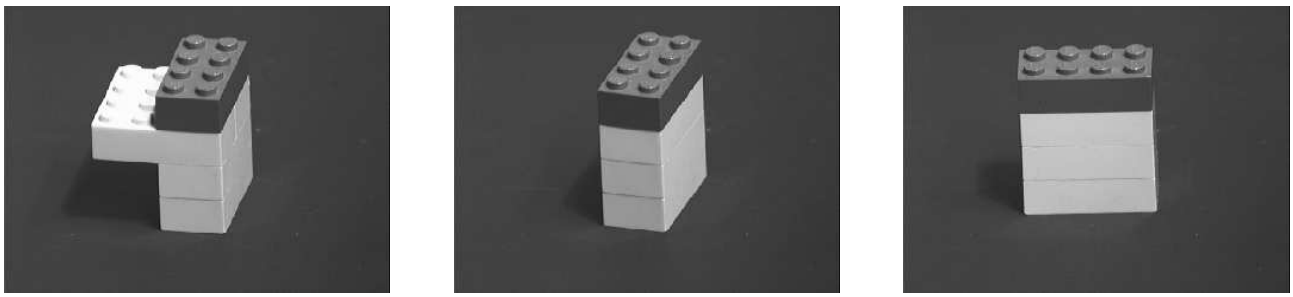

Figure 5: Two different 3–D objects which share a common view

increase the reliability of estimated pose parameters. Both the reliability of estimated pose parameters and the handling of mentioned ambiguities will be improved by using multiple views.

8 Mutual Information and Viewpoint Planning

If a calibrated camera is available, which is controlled by a robot, multiple views can be taken of a scene. The transforms between single views are known, because the extrinsic camera parameters can be computed using the robot's position. Let us assume, we have V views. Thus, we have the images $\boldsymbol{f}_1$, $\boldsymbol{f}_2, \ldots,$ $\boldsymbol{f}_V$, and the rotation and translation of the camera between single views. The transforms are denoted by $\{\Delta\boldsymbol{R}_v, \Delta\boldsymbol{t}_v; 1 \leq v \leq V\}$, wherein $\Delta\boldsymbol{R}_v$ and $\Delta\boldsymbol{t}_v$ symbolize the transform of the camera for the v–th view with respect to a reference position. For simplicity, we assume that the transform associated with the first view is the identity. If pose parameters $\boldsymbol{R}$ and $\boldsymbol{t}$ of an object are unknown and if multiple views are available, the position and orientation can be computed by solving the maximum likelihood estimation problem

$$\{\widehat{\boldsymbol{R}}, \widehat{\boldsymbol{t}}\} \quad = \quad \underset{\boldsymbol{R},\boldsymbol{t}}{\operatorname{argmax}} \prod_{v=1}^{V} p(\boldsymbol{X}|\boldsymbol{f}_v, \boldsymbol{a}, \boldsymbol{R}, \boldsymbol{t}, \Delta\boldsymbol{R}_v, \Delta\boldsymbol{t}_v) \quad . \tag{11}$$

Different views for recognition and pose estimation purposes can be selected randomly or following some heuristics or applying some optimality criterions. If a set of features is given, standard pattern recognition theory provides a various methods to select n–best features with respect to the classification task. Among them are algorithms based on information theory. The idea of these methods is to select those subsets with the highest amount of information. Since each view results in sensor data (or features), the selection of viewpoints is closely related to the problem of choosing the most discriminating elements of a, in principle infinite, set of features.

According to optimal viewpoint planning strategies, it is necessary at every moment to choose and to observe such images from the given object supplying the largest amount of information, i.e., eliminating the largest degree of uncertainty. The term *information* was mathematically formalized by Shannon, and the presented statistical characterizations of objects of previous sections allow the definition of the viewpoint selection problem in the sense of Shannon's information theory [6, 19].

For that purpose, we consider the process of generating images of objects as a transmission of information through a communication channel. We put an object of class Ω_κ into the communication channel. The output is the observable image $\boldsymbol{f}$. The input alphabet Ω of the channel is defined by the object classes $\Omega_1, \Omega_2, \ldots, \Omega_K$ and the output alphabet $\boldsymbol{F}$ results from the available images $\boldsymbol{f}$. Figure 6 illustrates different types of information, as they appear with respect to object recognition problems. Irrelevance summarizes the information which is added by the channel. This might be background features or some noise effects. The equivocation is the part of information which gets lost during the transmission through the channel. Examples are the depth data, which are eliminated by the projection to the image plane. The mutual information defines the rate of transmission of information through the channel related to the

input object. Therefore, mutual information is the most important measure of the assumed transmission channel. In contrast to standard definitions, mutual information depends here also on transformation parameters. These parameters are, however, not part of the observation.

The mutual information $I(\Omega_\kappa, \boldsymbol{f}; \boldsymbol{R}, \boldsymbol{t})$ (with integrated transform denoted by $\boldsymbol{R}$ and $\boldsymbol{t}$) of the communication channel is defined by

$$\begin{aligned} I(\Omega_\kappa, \boldsymbol{f}; \boldsymbol{R}, \boldsymbol{t}) &= \log \frac{p(\boldsymbol{f}, \Omega_\kappa; \boldsymbol{R}, \boldsymbol{t})}{p(\boldsymbol{f}; \boldsymbol{R}, \boldsymbol{t}) p(\Omega_\kappa; \boldsymbol{R}, \boldsymbol{t})} \\ &= \log \frac{p(\boldsymbol{f}, \Omega_\kappa; \boldsymbol{R}, \boldsymbol{t})}{\sum_{\Omega_\kappa} p(\boldsymbol{f}, \Omega_\kappa; \boldsymbol{R}, \boldsymbol{t}) p(\Omega_\kappa; \boldsymbol{R}, \boldsymbol{t})} \quad . \end{aligned} \tag{12}$$

Due to the statistical nature of model densities introduced in this paper, the probabilities required for the evaluation of mutual information can be evaluated, if the pose parameters are known. Since the position and orientation of objects influence the amount of information, we can give a formal definition of the *best* viewing direction:

$$\{\boldsymbol{R}_\kappa, \boldsymbol{t}_\kappa\} = \underset{\boldsymbol{R}, \boldsymbol{t}}{\operatorname{argmax}} \; I(\Omega_\kappa, \boldsymbol{f}; \boldsymbol{R}, \boldsymbol{t}) \quad . \tag{13}$$

With this information–theoretical formalization, the best views of objects with respect to a given model database are well–defined and can be computed.

For the practical use of this concept, we compute in an off–line pre–processing stage those pose parameters $\boldsymbol{R}_\kappa$ and $\boldsymbol{t}_\kappa$, which show the highest mutual information for each object class Ω_κ. This can be done by using the same global optimization methods which are also applied for pose estimation, and has to be recomputed, if the model database is extended for additional objects. If a set of most discriminating viewpoints is required, a sequence of best rotation and translation parameters with decreasing information can be computed using mutual information. Results of

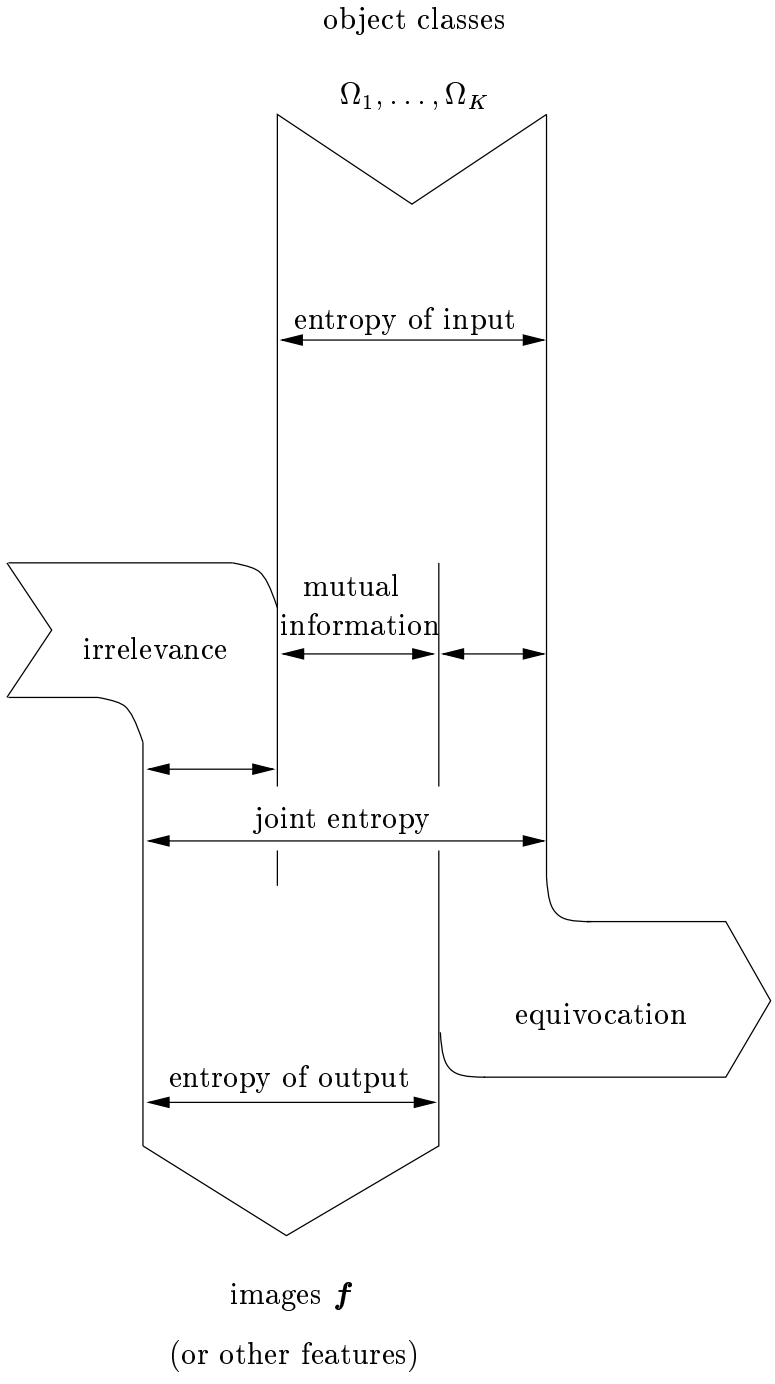

Figure 6: Illustration of relations between object recognition and mutual information

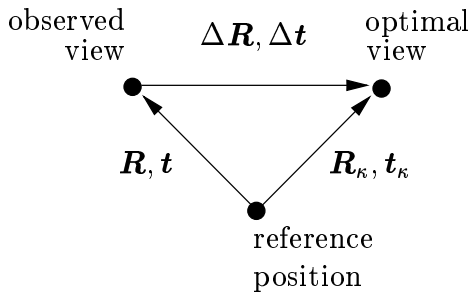

Figure 7: Computation of the transform required to find the best viewpoint

.

this pre–processing step can be used to implement an entropic active object recognition system, which gives the so–called most rational algorithm for recognition [6, p. 355].

Example: *Let us assume we observe an image* $\boldsymbol{f}$ *showing an object of an unknown class* Ω_κ. *The pose parameters can be estimated using this single view. If the most significant viewing direction (defined by* $\boldsymbol{R}_\kappa$ *and* $\boldsymbol{t}_\kappa$*) for object* Ω_κ *is known, we can compute the transformation parameters,* $\Delta\boldsymbol{R}$ *and* $\Delta\boldsymbol{t}$*, for the camera to the most discriminating viewpoint. If an initial estimate of* $\boldsymbol{R}$ *and* $\boldsymbol{t}$ *is computed by (9) the transformation can be computed using the graph shown in Figure 7.* □

9 Experimental Results

The experimental evaluation considers recognition experiments using both gray–level images and point features for classification purposes.[1]

[1] In case of acceptance, the final paper will include more experiments. The work is in progress.

Figure 8: Objects of the model data base

9.1 Object Recognition using Intensity Images

In the first set of experiments, we avoid segmentation and reduce gray–level images to lower dimensional features vectors by a Karhunen–Loéve transforms [15]. The spatial distribution of these vectors is modelled by single Gaussians. For model generation purposes, there are 100 training views of each object available. The disjoint test set contains also 100 views of each object. The considered four object classes are shown in Figure 8. The computed recognition rate with homogeneous background is 50 %, if 20–dimensional vectors are used. If we use the absolute values of the 2–D Fourier transform instead of intensity values, the overall recognition rate increases to 100 % for the given test set. The statistical modeling of transformed gray–level images by mixtures of spatial distributions, here, leads to a recognition rate of 91 %.

9.2 Object Recognition using Point Features

The reduction of intensity images to point features leads to the expected decrease of recognition rates. Instead of 100 % using Fourier transforms, only 82 % of the objects are classified correctly. Nevertheless, point features are worth being considered for further experiments. Corners of polyhedral objects allow the estimation of higher dimensional mixture densities, even if only projected training data are available. This is not the case, if only intensity are considered without having a suitable illumination model. Assume we model the 3–D corners of

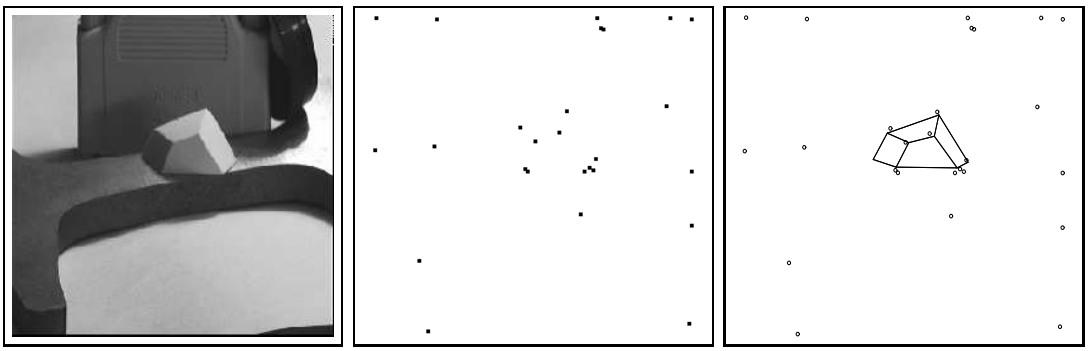

Figure 9: Maximum likelihood estimate of pose parameters resulting from point features

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

an object using a mixture of Gaussians. The 3–D mean vectors and the covariance matrices, for instance, have to be estimated using 2–D, orthographic projections from several views. Due to the hidden depth information, the EM algorithm is applied for model generation purposes. For pose estimation, a the solution of a global optimization task is required. The global optimization is done by probabilistic search methods. A reliable global optimization of the objective function within the five–dimensional search space requires in average four hours on a HP 735 (125 MIPS), if twelve mixture components are involved. The use of marginals to reduce the complexity of global optimization, induces a descend of the computation time to 1.5 minutes. The correct pose parameters on synthetic data (400 2–D images) could be found with a probability of 82 %. The experiments also show, the marginals do not only reduce the dimensions of the search space, but also the time for evaluating density functions. An example for maximum likelihood estimates of pose parameters can be found in Figure 9. The implementation of a parallel search method which distributes the optimization problem to several processors leads to speed–ups summarized in Table 1.

Recognition experiments on 1600 images showing 3–D objects with homogeneous back-

ground result in classification rates of 70 % if single views are used for identification.

9.3 Object Recognition using Multiple Views

The use of multiple views for object recognition shows two advantages: on the one hand the objective function for pose estimation shows a unique global maximum if additional views are considered, on the other hand the discriminating power increases. The use of multiple views for pose estimation shows remarkable improvements regarding the correct pose parameters. We run experiments using 400 random views of synthetic objects and the correct pose parameters increased from 96 % to 100 % if a second view is used. Existing ambiguities 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.

10 Summary and Conclusions

This paper has shown a Bayesian approach to object recognition using gray–level images or induced geometrical features like points or lines. The statistical modeling of objects allows the use of the Bayesian decision rule, and thus the implementation of a theoretically optimal classifier is possible. Object models correspond to density functions, wherein mixtures and density transforms are important. In this context, model generation procedures as well as pose computation are related to parameter estimation problems. For computational efficiency, we use marginals which reduce the dimension of the search spaces for several applications. Another important concept deals with problems of viewpoint planning and priors of viewing directions based on mutual information with integrated feature transform. This information theoretical formalization results in an entropic object recognition system, which uses sensor

data with the highest amount of information. The best views are generated maximizing the mutual information between object classes and observed images.

The introduced statistical framework is suitable for object recognition, pose computation, and viewpoint planning. Nevertheless, two important issues remain unexplored:

- How many views and how many features are required for a certain recognition rate?
- Which resolution and which type of features are optimal with respect to the given computer vision problem?

Acknowledgements

The authors like to thank Dr. W. Wells III (AI lab, MIT) and Bernt Schiele (IMAG–GRAVIR, I.N.P., Grenoble) for fruitful discussions and suggestions.

11 References

1. T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley Publications in Statistics. John Wiley & Sons, Inc., New York, 1958.
2. D.H. Ballard and C.M. Brown. *Computer Vision.* Prentice-Hall, Englewood Cliffs, NJ, 1982.
3. P. A. Devijver and M. Dekesel. Learning the parameters of a hidden Markov random field image model: A simple example. In P. A. Devijver and J. Kittler, editors, *Pattern Recognition Theory and Applications*, volume 30 of *NATO ASI Series F: Computer and System Sciences*, pages 141–163. Springer, Heidelberg, 1987.
4. K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, Boston, 1990.
5. S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984.
6. S. Guiasu. *Information Theory with Applications.* McGraw–Hill, New York, 1977.
7. J. Hornegger and H. Niemann. Statistical learning, localization, and identification of objects. In ICCV 95 [9], pages 914–919.
8. X.D. Huang, Y. Ariki, and M.A. Jack. *Hidden Markov Models for Speech Recognition.* Edinburgh University Press, 1990.
9. *Proceedings of the* 5^{th} *International Conference on Computer Vision (ICCV)*, Boston, June 1995. IEEE Computer Society Press.
10. A. K. Jain and P. J. Flynn, editors. *Three–Dimensional Object Recognition Systems*, Amsterdam, 1993. Elsevier.
11. S. Z. Li. *Markov Random Field Modeling in Computer Vision.* Springer, Heidelberg, 1996.
12. S. Z. Li. Parameter estimation for optimal object recognition: Theory and application. *International Journal of Computer Vision*, 21(3):1–17, March 1997.

13. R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, Inc., New York, 1987.
14. J. L. Mundy, A. Zisserman, and D. Forsyth, editors. *Application of Invariance in Computer Vision*, volume 825 of *Lecture Notes in Computer Science*, Heidelberg, 1994. Springer.
15. H. Murase and S. K. Nayar. Visual learning and recognition of 3–D objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, January 1995.
16. J. Ponce, Zisserman, and M. Hebert, editors. *Object Representation in Computer Vision*, volume 1144 of *Lecture Notes in Computer Science*, Heidelberg, 1996. Springer.
17. L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993.
18. I. Rigoutsos. Massively parallel Bayesian object recognition. PhD thesis, Courant Institute of Mathematical Sciences, New York, 1992.
19. B. Schiele and J.L. Crowley. Transinformation of object recognition and its application to viewpoit planning. *Robotics and Autonomous Systems*, http://pandora.imag.fr/Prima/schiele/home.html,to appear, 1997.
20. I. Shimshoni. Interpreting images of polyhedral objects in the presence of uncertainty. PhD thesis, Department of Computer Science, University of Illinois, Urbana–Champaign, 1995.
21. M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, November 1991.
22. M. A. Tanner. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* Springer Series in Statistics. Springer, Heidelberg, 1993.
23. K. A. Tarabanis, P. K. Allen, and R. Tsai. A survey of sensor planning in computer vision. *IEEE Transactions on Robotics and Automation*, 11(1):86–104, February 1995.
24. F. C. D. Tsai. Geometric hashing with line features. *Pattern Recognition*, 27(3):377–391, 1994.
25. J. K. Tsotsos and Y. Ye. Sensor planning in 3d object search. In *Proceedings of the 4th International Symposium on Intelligent Robotic Systems*, pages 131–136. IEEE Computer Society Press, July 1996.
26. P. Viola and W. M. Wells III. Alignment by maximization of mutual information. In ICCV 95 [9], pages 16–23.
27. D. Weinshall and M. Werman. On view likelihood and stability. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):97–109, January 1997.
28. W. M. Wells III. Statistical Object Recognition. PhD thesis, MIT, Department of Electrical Engineering and Computer Science, Massachusetts, February 1993.
29. S. Yakowitz. Unsupervised learning and the identification of finite mixtures. *IEEE Transactions on Information Theory*, 16:330–338, 1970.