M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 453
accepted for publication in IJCV, International Journal on Computer Vision

Recognition without Correspondence using Multidimensional Receptive Field Histograms

Bernt Schiele and James L. Crowley

The Media Laboratory, Room E15-384a,
Massachusetts Institute of Technology
20 Ames St., Cambridge, MA 02139
email: bernt@media.mit.edu

Abstract

The appearance of an object is composed of local structure. This local structure can be described and characterized by a vector of local features measured by local operators such as Gaussian derivatives or Gabor filters. This article presents a technique where appearances of objects are represented by the joint statistics of such local neighborhood operators. As such, this represents a new class of appearance based techniques for computer vision. Based on joint statistics, the paper develops techniques for the identification of multiple objects at arbitrary positions and orientations in a cluttered scene. Experiments show that these techniques can identify over 100 objects in the presence of major occlusions. Most remarkably, the techniques have low complexity and therefore run in real-time.

1 Introduction

The paper proposes a framework for the statistical representation of the appearance of arbitrary 3D objects. This representation consists of a probability density function or joint statistics of local appearance as measured by a vector of robust local shape descriptors. The object representations are acquired automatically (learned) from sample images. Multidimensional histograms are introduced as a practical and reliable means for the approximation of the probability density function for local appearance. An important result of this paper is that the representation based on joint statistics of local neighborhood operators provides a reliable means for the representation and recognition of large sets of objects (over 100 objects) at arbitrary 3D positions and orientations in cluttered scenes.

Three different recognition algorithms are proposed within this framework and evaluated experimentally. The first algorithm compares the probability distribution of local neighborhood operators of a test image to the distributions of learned objects. Recognition is achieved by applying statistical divergence measurements which can be seen as a generalization of the color indexing scheme of Swain and Ballard [Swain and Ballard, 1991]. The second recognition algorithm calculates probabilities for the presence of objects based on a small number of vectors of local neighborhood operators. The experiments demonstrate that in the typical case, a small number of vectors is sufficient to obtain the correct object hypothesis from a database of 100 objects. In particular, experimental results show the robustness of the approach to partial occlusion. The most remarkable property of the the algorithm is that it relies on neither the calculation of correspondence nor figure ground segmentation of the object in the scene.

The second algorithm is extended to recognize multiple objects in cluttered scenes by using *local appearance hashing*. The capacity of the algorithm to recognize objects in cluttered scenes without relying on the calculation of correspondence is demonstrated experimentally. Due to its low complexity this algorithms runs on a standard Silicon Graphics O2-machine at 10Hz using the OpenGL-extension for real-time convolution of images.

It has been shown that the segmentation problem has exponential complexity[1] in the size of the image considering no knowledge about the scene and in particular assuming no knowledge about which objects might be in the scene [Tsotsos, 1989]. However, the task-oriented visual search as e.g. in the case of segmenting objects knowing which objects are in the scene, has only linear complexity. The probabilistic algorithm of section 6 and its extension in section 8 calculate object hypotheses with linear complexity (in the number of used image measurements and number of objects). This low complexity is mostly due to the fact that no correspondence and no segmentation are calculated. In that sense, this

[1] More specifically the bottom-up visual search task as defined in [Tsotsos, 1989] is NP-complete in the size of the image

paper proposes algorithms with linear complexity in order to obtain object hypotheses which can be used subsequently by a segmentation algorithm with linear complexity.

The next section briefly discusses closely related object recognition work. Since we use Gaussian derivatives throughout this paper we introduce them in section 3. Section 4 derives a general *statistical object representation framework* based on the statistics of local neighborhood operators. Section 5 introduces histogram matching as the first algorithm for the recognition of objects. Even though histogram matching enables the discrimination between 100 objects, the nature of the approach is global. Section 6 therefore proposes a local recognition algorithm which calculates probabilities of objects based on a small number of vectors of neighborhood operators. The comparison of experimental results shows that this algorithm is highly robust to partial occlusion. This enables us to define in section 8 an algorithm based on *local appearance hashing* which is particularly suited for the recognition of multiple objects in cluttered scenes.

2 Related object recognition work

This section briefly discusses closely related object recognition work (see [Object representation, 1996] [Pope, 1995] [Grimson and Huttenlocher, editors, 1992] [Grimson and Huttenlocher, editors, 1991] for more comprehensive reviews).

2.1 Histogram based approaches

Swain and Ballard [Swain and Ballard, 1991] have proposed to represent an object by its color histogram (approximating its color distribution). Objects are identified by matching a color histogram from an image region with a color histogram from a sample of the object. Their technique has been shown to be remarkably robust to changes in the object's orientation, changes of the scale of the object, partial occlusion or changes of the viewing position. Even changes in the shape of an object do not necessarily degrade the performance of their method. The robustness to scale and rotation are mainly provided by the use of color. The robustness to changes in viewing angle and to partial occlusion are due to the use to *histogram matching*. However, the major drawback of their method is its sensitivity to lighting conditions such as the color and the intensity of the light source. Also, many object classes cannot be described by color alone.

In order to reduce the sensitivity to illumination intensity changes several authors have introduced color invariances. [Healey and Slater, 1994] for example calculate moment invariants of the entire color histogram (assuming a constant intensity change over the entire image). [Funt and Finlayson, 1995] use derivatives of the logarithms of the color channels in order to provide illumination invariant features (assuming a locally constant illumination). More recently [Finlayson et al., 1998] introduced a color image normalization which is invariant to light intensity and light color changes. Another interesting extension [Ennesser and Medioni, 1995] uses local color histograms of the test image in order to deal with more cluttered scenes.

Since not all objects can be described and recognized by color alone, color histograms have been combined with geometric information (e.g. [Slater and Healey, 1995] [Matas et al., 1995]). In particular, the SEEMORE-system [Mel, 1997] uses 102 different feature channels which are each sub-sampled and summed over a pre-segmented image region. The 102 channels compromise color, intensity, corner, contour shape and Gabor-derived texture features. Strikingly good experimental results are given on a database of 100 pre-segmented objects of various types. Most interestingly, a certain ability to generalize outside the database has been observed.

The color histogram approach is an attractive method for object recognition, because of its simplicity, speed and robustness. Many image retrieval system use color histograms among other cues (e.g. [Flickner et al., 1995] [Belongie et al., 1998]) which is motivated by the fact that many images contain characteristic colors. Since many objects cannot be described by color alone this paper generalizes the color histogram approach to *multidimensional receptive field histograms*. Such receptive fields may capture local structure, shape or any other local characteristic appropriate to describe the local appearance of an object.

2.2 Object recognition based on local characteristics

[Lamdan and Wolfson, 1988] introduces *geometric hashing* as a general framework for recognizing overlapping and partially occluded objects. Object models consist of sets of interest points. The representation of the sets is made invariant to an affine transformation by using three points as an affine basis[2]. In order to reduce the calculation time and the complexity of recognition all possible triplets of interest points are used as basis and the coordinates of the remaining interest points are stored in a hashtable. During recognition sets of interest points are extracted from

[2]In the case of a projective transformation a four point basis is used. For similarity transformations only two points are needed[Lamdan and Wolfson, 1988].

the scene and used for indexing into the hashtable and voting for object models. Recognition therefore becomes a point matching task. [Grimson et al., 1994] provide a theoretical analysis of the sensitivity of geometric hashing. The main result is that the probability of false positives (during voting) increases considerably in the presence of moderate noise in the data points. An improved probabilistic voting scheme addresses this issue [Rigoutsos and Hummel, 1993].

The robustness and the repeatability of the interest point detector in the presence of affine transformations is crucial [Schmid et al., 1998]. By using only point features the algorithm may result in a large number of false positives. Therefore, [Lamdan et al., 1988] [Wolfson, 1990] use not only interest points but also other features. However, the feature choice is limited since they require invariance to affine transformations.

[Ballard and Wixson, 1993] and [Rao and Ballard, 1995] propose to represent objects (or object patches) by a high–dimensional "iconic" feature vector. Such high-dimensional object representations have the favorable property that they can be subjected to considerable noise before they are confused with the vectorial representation of other objects. More specifically, the feature vector includes 45 responses of nine oriented Gaussian filters at five different scales ($9 \times 5 = 45$). Using the steerability of Gaussian derivatives [Freeman and Adelson, 1991], the feature vector is made rotational invariant. During training and object recognition a figure ground segmentation is performed and the vectors are stored in a generalized version of Karneva's sparse distributed memory.

One drawback of the proposed feature vector is its relatively large support (about 128×128 pixels[3]) which makes the approach sensitive to occlusion. Reducing the support of the feature vector would compromise on the uniqueness of the filter response. [Ballard and Rao, 1994] introduce a separate algorithm which can account for partial occlusions. The basic idea is to reconstruct an image patch approximately by a pseudo inverse transformation from a single feature vector. By masking the occluded parts the reconstructed image can be compared with the observation in the image.

[Rao and Ballard, 1997] propose a predictive Kalman filter hierarchy which combines input-driven bottom-up signals with the expectation-driven top-down signals. This architecture can be seen as a hierarchy of local representations which are learned simultaneously. It is used to implement a dynamic recognition algorithm using pattern completion during occlusions. The hierarchy is used to explain neural responses of a monkey freely viewing a natural scene.

A reliable object recognition algorithm has been proposed in [Schmid and Mohr, 1997]. Each interest point in an image is described by a nine–dimensional rotational invariant vector of local characteristics based on Gaussian derivatives, originally proposed in [Koenderink and Doorn, 1987]. Finally, the vector responses of all interest points of an image are stored in a hash table indexed by the nine–dimensional vector. In this sense the approach is a synthesis of the two previous ones: local representation by a hash table and rich description of local structure by a vector of local characteristics.

The principal application of the approach is the correspondence problem between a test image and the stored images in the hash table. In addition, the approach is suitable for object (or image) recognition which can seen as a correspondence problem. By applying the interest point detector to a test image and by calculating the vector responses for the interest points the algorithm votes for different images (or objects). The voting technique is made more selective by combining the vector responses with geometrical invariants between different interest points . Another possibility for improvement is the use of a probabilistic voting scheme [Mohr et al., 1997].

Impressive experimental results have been presented on a database of several hundred objects. Nevertheless, arguably the weakest point of the approach is the application of an interest point detector [Schmid et al., 1998]. The success of the approach relies on the repeatability of the interest point detector over different images and viewing conditions which is difficult to achieve, particularly in unconstrained environments

2.3 Eigenvector approaches

Recently many researchers [Sirovich and Kirby, 1987] [Turk and Pentland, 1991] [Murase and Nayar, 1995] [Moghaddam and Pentland, 1995] [Ohba and Ikeuchi, 1996] have used the Karhunen–Loeve transformation [Fukunaga, 1990] for the calculation of *eigenpictures* in the context of object recognition. The main advantage of this approach is the representation of each image by a small number of coefficients, which can be stored and searched efficiently. Even though very successful, the approach has two major drawbacks: the first drawback is due to the fact that any change of individual pixel values, caused for exam-

[3] The responses of the 45-dimensional feature vector are calculated with a 8×8 kernel at five different levels of an image pyramid. Since each level of the image pyramid is reduced by a factor of 2, the overall support of a single vector is in the order of 128×128 pixels. The vector therefore cannot be called local, since it already covers a $\frac{1}{16}$ of a typical 512×512 image.

ple by translation, by scale change, by image plane rotation or by illumination changes, will change the eigenvector representation of an image.

Two principal possibilities exist in order to deal with this difficulty: either each image is normalized prior to projection onto the eigenspace or the eigenspace is calculated under consideration of all possible changes. Even though a powerful normalization function can be implemented in the special case of face recognition it is difficult to assume such a function in the general case of 3D–object recognition. In the general case of 3D–object recognition a pre-segmentation step is assumed prior to the projection onto the eigenspace [Murase and Nayar, 1995]. The second major drawback of the approach is that the modeling of each image is global, which makes the approach sensitive to partial occlusion.

3 Vector of local neighborhood operators

Measurements of local object appearance can be obtained by a multi–dimensional vector of local neighborhood operators. The neighborhood operators which we employ below are not restricted to a particular family of objects nor does the approach rely on the use of a particular set of features. Nevertheless, it is necessary to formulate minimal requirements. The first requirement is the *locality* of the features. As we have shortly mentioned in section 2, global features are sensitive to partial occlusion as well as local image disturbances such as specular reflections. The second requirement concerns the *sensitivity* of the features. We can list three categories concerning the sensitivity of features:

invariance: invariant features are considered constant with respect to certain transformations (such as affine and projective transformations),

equivariance: the values of equivariant features are a function of a certain transformation,

robustness: the values of robust features change slowly in the presence of certain transformations. Such features are often called quasi-invariant.

The invariance of features is the most powerful property yet the most difficult to obtain in reality. Whenever possible we should use invariant features. Unfortunately invariant features typically impose unacceptable restrictions on the set of object classes which can be recognized. Furthermore, most invariant local features are based on the calculation of higher order derivatives and thus create practical problems related to instability, as well as locality problems. Either of these constraints would limit the generality of our approach. Consequently, we find it necessary to relax the requirement of invariance.

Equivariant features vary as a function of a certain transformation. An example is the equivariant property of Gaussian derivatives with respect to image plane rotations and scale changes. Unfortunately, equivariance is restricted to certain classes of image structure, and can not be obtained in a general manner.

In general, robustness or quasi–invariance can be attained more easily. Robust features will change slowly and in a predictable manner with respect to changes of the object's appearance. Many local features exist which are robust to appearance changes such as viewing position, illumination and scale. In our experiments, we only employ features which can be calculated locally and which are robust with respect to image noise, blur, image plane rotation and scale.

Section 3.1 introduces Gaussian derivatives, their steerability with respect to image plane rotation and the equivariance property to scale change. Gaussian derivatives are widely used in computer vision. Their popularity is due to their generality (eigenpictures of large numbers of image patches resemble derivatives of Gaussians [Rao and Ballard, 1995]), their capacity to model the response of neural cells [Young, 1986] and the existence of a recursive implementation [Deriche, 1987]. Furthermore, Gaussian derivatives (as well as Gabor filters) are robust to scale changes of approximately ±20% [Schmid and Mohr, 1997].

Gabor filters [Gabor, 1946] [Westelius, 1992] [Daugman, 1993] satisfy the same constraints as Gaussian filters (robustness, steerability to image plane rotation, equivariance to scale changes). During earlier experiments (not reported below) Gabor filters obtained almost identical results as Gaussian derivatives. Even though color has not been used in our experiments, invariant color descriptors [Nagao, 1995] [Funt and Finlayson, 1995] provide a natural extension of the proposed statistical object representation technique described below. One can also consider the use of texture features [Haralick, 1979] [Mao and Jain, 1992] or low–level geometric features and perceptual significant groups thereof [Pope and Lowe, 1996].

3.1 Gaussian derivatives

Gaussian derivatives are widely used in the literature and well understood [Freeman and Adelson, 1991] [Rao and Ballard, 1995]. By using Gaussian derivatives we can explicitly select the scale. Additionally, we can "steer" the derivative to arbitrary orientations: it is possible to calculate the nth order Gaussian derivative of the orientation ϕ based on a linear combination of a finite number of nth order derivatives. This section describes Gaussian derivatives in general, develops the equivariance property to scale and finally summarizes the "steerability" to image plane rotation.

Given the Gaussian distribution $G^\sigma(x,y)$:

$$G^\sigma(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

The nth order Gaussian derivative in direction $\vec{v} = (\cos\phi \ \sin\phi)^T$ is defined by:

$$G^\sigma_{n,\phi}(x,y) = \frac{\partial^n}{\partial \vec{v}^n} G^\sigma(x,y) \tag{2}$$

In this article we use Gaussian derivatives up to the second order. Therefore, we will introduce a particular notation for the derivatives used. We define the x–axis parallel to the vector $\vec{v} = (1 \ 0)^T$, which corresponds to $\phi = 0°$. The y–axis is defined by $\phi = 90°$ and is therefore parallel to $\vec{v} = (0 \ 1)^T$. The derivatives in x– and y–direction are given by:

$$\begin{aligned} G^\sigma_x(x,y) &= G^\sigma_{1,0°}(x,y) = -\frac{x}{\sigma^2} G^\sigma(x,y) && (3)\\ G^\sigma_y(x,y) &= G^\sigma_{1,90°}(x,y) = -\frac{y}{\sigma^2} G^\sigma(x,y) && (4) \end{aligned}$$

Based on these first order derivatives we can define the magnitude $Mag(x,y)$ of the first Gaussian derivative:

$$Mag(x,y) = \sqrt{(G^\sigma_x(x,y))^2 + (G^\sigma_y(x,y))^2} \tag{5}$$

Based on two second order derivatives $G^\sigma_{xx}(x,y)$ and $G^\sigma_{yy}(x,y)$ the well known Laplace operator $Lap(x,y)$ can be defined:

$$\begin{aligned} G^\sigma_{xx}(x,y) &= (\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2}) G^\sigma(x,y) && (6)\\ G^\sigma_{yy}(x,y) &= (\frac{y^2}{\sigma^4} - \frac{1}{\sigma^2}) G^\sigma(x,y) && (7)\\ Lap(x,y) &= G^\sigma_{xx}(x,y) + G^\sigma_{yy}(x,y) && (8) \end{aligned}$$

3.2 Equivariance of Gaussian derivatives to scale

As mentioned above local neighborhood operators should be calculated at an particular scale. Given a two–dimensional function $p(x,y)$ and its scaled version $f(x,y) = p(sx,sy)$ analysis tells us:

$$\begin{aligned} f(x,y) &= p(sx,sy) && (9)\\ \frac{\partial}{\partial x} f(x,y) &= s\frac{\partial}{\partial x} p(sx,sy) && (10)\\ &\vdots \\ \frac{\partial^n}{\partial x^n} f(x,y) &= s^n \frac{\partial^n}{\partial x^n} p(sx,sy) && (11) \end{aligned}$$

Following the above equations, the nth order derivative of the function f can be calculated on the basis of the nth order derivative of $p(sx,sy)$. This calculation assumes exact knowledge of the function p. In computer vision the exact knowledge of p cannot in general be assumed. By using Gaussian derivatives the nth order derivative of $p(sx,sy)$ can be calculated based on $p(x,y)$. In the following we show this property for the first order derivative. We define the first order derivative of f as:

$$\frac{\partial}{\partial x} f(x,y) = G^\sigma_x(x,y) \star f(x,y) \tag{12}$$

where $G^\sigma_x(x,y)$ is the Gaussian derivative (see equation 3) and $\star$ is the convolution operator. Therefore we obtain (together with equation 10):

$$\begin{aligned} \frac{\partial}{\partial x} f(x,y) &= s\frac{\partial}{\partial x} p(sx,sy) && (13)\\ &= sG^\sigma_x(x,y) \star p(sx,sy) && (14)\\ &= sG^{\sigma s}_x(x,y) \star p(x,y) && (15) \end{aligned}$$

The equation shows that we can calculate the first order derivative of f on the basis of the first order derivative of $p(x,y)$, which we call the *adaptation of the Gaussian derivative to scale*. In a similar way we obtain an equation for the adaptation of the nth order derivative to scale:

$$\frac{\partial^n}{\partial x^n} f(x,y) = s^n G^{\sigma s}_{x^n}(x,y) \star p(x,y) \tag{16}$$

Following this equation, we can calculate the nth order derivative of a function $f(x,y)$ directly based on function $p(x,y)$ (when f is a scaled version of p: $f(x,y) = p(sx,sy)$). In order to employ this property the scale factor s must be known, which cannot in general be assumed. Usually we calculate the derivative for different values of s. Additionally, the support for the calculation of the nth order derivative of p has to be adapted. This is expressed by the adaptation of the standard deviation σs of the Gaussian filter.

We call the adaptation of the Gaussian derivatives to scale changes by the factor s the *equivariance* property of the Gaussian derivatives to scale. As expected, the *equivariance to scale* is not only true for neighborhood operators based on Gaussian derivatives. The same property holds, for example, for Gabor filters due to their Gaussian envelope.

3.3 Steerability of Gaussian derivatives to image plane rotation

In order to calculate the filter response (for example for a Gaussian filter) at an arbitrary orientation ϕ the corresponding version of the filter can be calculated. If the orientation is not known beforehand or if a particular filter response has to be calculated for many different orientations, it is desirable to define a finite set of basis filters and an interpolation rule, which allows the calculation of the filter response based only on the response of the basis set. [Freeman and Adelson, 1991] show that the minimal number of interpolation functions for the nth order Gaussian derivative is $n+1$. This correspond e.g. to the well known interpolation rule for the first order Gaussian derivative:

$$G^{\sigma}_{1,\phi} = \cos\phi G^{\sigma}_{x} + \sin\phi G^{\sigma}_{y} \quad (17)$$

4 Statistical object representation

The appearance of an object is composed of local structure. This local structure can be described and characterized by a vector of local neighborhood operators. We propose to represent 3D objects by the joint statistics of local structure, which can be calculated reliably from sample images of the objects. The probability function of an object and therefore the object's model is learned automatically.

Let's assume we have chosen a fixed measurement set $M = \cup_k m_k$ composed of vectors m_k of local neighborhood operators. The probability density function over the measurement set M for a certain object o_n varies with the changes of the appearance of the object which should be modeled within the probability density function. Five categories of possible changes can be listed (see figure 1):

Similarity transformation: three translational degrees of freedom (t_x, t_y and t_z) and one rotational degree of freedom (r_z) can be identified (see figure 1).

3D transformation of the object: two rotational degrees of freedom (r_x and r_y) exist in addition to the similarity transformation (see figure 1).

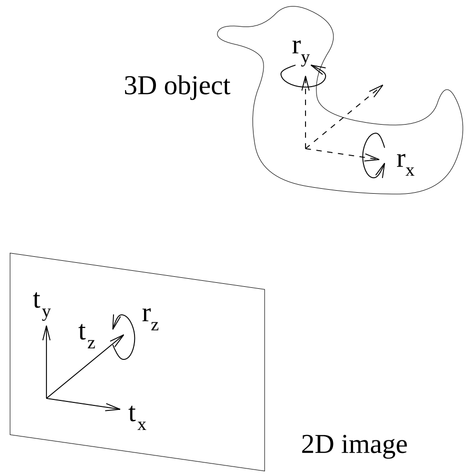

Figure 1: Different components of rotation and translation of a 3D object

Scene changes: this includes partial occlusion and background change.

Light conditions: this includes changes in the intensity, color and direction of the light source.

Imaging conditions: different types of signal disturbance as signal noise, quantization error and blur.

By writing the probability density function of the object o_n, parameterized by variables of these changes, we obtain:

$$p(M|o_n, R, T, S, L, I) \quad (18)$$

where M is the set of local image measurements m_k, o_n is the label of an object (or object class), R describes the three rotational degrees of freedom, T the three translational degrees of freedom, S the scene changes, L the light changes and I the imaging conditions.

In general it is difficult to obtain a reliable estimate of such a high–dimensional probability density function. The difficulty is due to the fact that the number of training examples is exponential in the number of dimensions of the density function [Intrator and Gold, 1993]. The most effective way to reduce the number of free parameters is to choose local image measurements which are invariant to different parameters. Such invariant properties are used by many researchers [Burkhardt and Zisserman, 1992]

[Mundy and Zisserman, 1992] [Mundy et al., 1993] and applied successfully in various ways. Unfortunately the obtained invariants are very restrictive to certain types of objects. Robust or quasi-invariant local image measurements are often an alternative since they are less restrictive than invariants and since we can typically identify a reasonable range of changes where their values are near constant.

One category of changes, the *imaging conditions*, is characterized by changes which cannot be controlled in general. In this case the approach relies on the fact that local descriptors can be calculated robustly with respect to such changes. The analysis of the robustness therefore demands special consideration. [Schiele, 1997] examines the robustness of local image measurements and different normalization techniques in the presence of different sources of noise. For the second category, the *light conditions*, exist many normalization techniques but none of them is satisfactory for the general case. Currently, we are using an energy normalization technique of the filter output which has shown to provide good results in the presence of different light condition changes.

Scene changes due to *partial occlusion* and *background change* are difficult to model. One possibility is to include partial occlusion and background change in the estimation process of the probability density function. [Hornegger and Niemann, 1995] propose to model partial occlusion as a particular object: the background. By introducing a probability for the background – which is directly related to the observed portion of the object – the probability of the presence of an object can be calculated. The recognition process therefore estimates not only the object's label and its pose but also the portion of occlusion. Recognition becomes an iterative optimization process, which is elegant but relatively time consuming. In contrast to this approach we propose in section 6 a probabilistic object recognition approach which is able to recognize objects by the observation of a small portion of the object. This algorithm makes the recognition process not only fast but also robust to partial occlusion. As a result we do not have to consider partial occlusion in the modeling of the probability density function of an object. In our context, background changes are considered as a special case of partial occlusion.

The correspondence problem between the object model and a test image is in general difficult and time consuming. In order to avoid this problem we do not represent the two translational parameters t_x and t_y in the probability density function. Several advantages motivate this choice: First of all and as just mentioned the translational correspondence problem does not exist. Secondly the estimation of the probability density functions becomes feasible. The estimation becomes feasible because of the dimensionality reduction of the density function and also because of the amount of training samples which is provided by images of an object. A typical 512×512 image of an object provides about $500^2 = 250,000$ training samples for the estimation of the probability density function of the object.

The third translational parameter t_z can be treated directly by the transformation of the image pattern. Throughout the article we employ the equivariance property of local descriptors to scale in order to account for t_z. The image plane rotation parameter r_z can be accounted for by using local descriptors, which are invariant to r_z. Such invariants have been used for example by [Schmid and Mohr, 1997]. The main disadvantage of these local descriptors is that rotational information is lost. Another disadvantage is the underlying assumption that all rotations are equally probable, which cannot in general be assumed. In the context of this work we use both image plane rotation invariant and variant local descriptors. In the case of variant descriptors, image plane rotation is managed by the rotational steerability of local descriptors.

The two rotational degrees of freedom r_x and r_y represent a *viewpoint change* of the observer. Several authors [Burns et al., 1990] [Clemens and Jacobs, 1991] show that the non–existence of viewpoint invariant descriptors for the general case. Nevertheless, useful descriptors exist in special cases [Mundy and Zisserman, 1992] [Mundy et al., 1993]. As mentioned earlier we do not want to restrict our approach to such specialized invariants. We model therefore the two parameters r_x and r_y in the probability density function.

What remains from the original probability density function (equation 18) are three components of the rotation and one component of the translation:

$$p(M|o_n, r_x, r_y, r_z, t_z) \tag{19}$$

By considering an L–dimensional vector m_k of local image measurements the statistical representation of an object o_n is given by an $L+4$–dimensional probability density function. In the case of image plane rotation invariant descriptors the representation is given by an $L+3$–dimensional probability density function.

4.1 Representation by multidimensional histograms

Different possibilities exist in order to estimate and represent the probability density function (equation 19) of an object. Typically, parametric and non-parametric estimation schemes can be distinguished. Parametric estimators assume a certain type of distribution as for example a poison distribution or a Gaussian distribution. The learning therefore becomes an estimation of the parameter of the assumed distribution. Hornegger and Niemann [Hornegger and Niemann, 1995] use parameterized mixtures of multivariate Gaussian distributions including a feature transform. Their statistical model consider the statistical behavior of features, feature matching, as well as the projection from the model into the image space. The assumption of a mixture of Gaussian distributions has been shown to be appropriate for point features but cannot be assumed for more general local image measurements.

The other principal possibility is a non–parametric estimator for the probability density function. In the context of high–dimensional density functions essentially two methods can be applied: histogramming and kernel function estimates [Popat and Picard, 1994]. The main advantage of histogramming is that the training samples are well represented. This property is desirable in our context since we aim to show that the proposed statistical object representation provides a reliable and discriminant means for the recognition of a large number of objects. This implies that the representation should preserve all information and in particular the discriminant information and therefore motivates the choice of histograms. On the other hand kernel functions typically allow the generalization from training samples. However, in our case the use of kernel functions only made a marginal difference with respect to generalization. This is mainly due to the fact that the number of training samples is sufficiently large in order to obtain a reliable estimate of the probability density function using histograms.

Consequently, we represent the probability density function of a certain object by several multidimensional histograms over the measurement set M. As an example figure 2 shows two-dimensional histograms of two different objects each corresponding to a particular viewpoint, image rotation and scale. The histogram of a particular viewpoint $(r_x\ r_y)$, at a particular image plane rotation r_z and at a certain scale t_z is given by:

$$H(M|o_n, r_x, r_y, r_z, t_z) \tag{20}$$

In order to obtain these histograms we have to take several images of the object. The number of training images can be reduced considerably by using the steerability to image plane rotation and the equivariance property of local image measurements to scale changes. The steerability and equivariance property of Gaussian derivatives is described in section 3. That implies that we can take a single image per viewpoint $(r_x\ r_y)$ and calculate several histograms which correspond to different image plane rotations r_z and scales t_z of the object.

The histograms of different viewpoints have to be estimated from several images of the object corresponding to several viewpoints (represented by r_x and r_y). [Schiele, 1997] examines the number of histograms which are needed for the representation of a 3D object. We concluded from experiments that a small number of histograms are sufficient in order to obtain high recognition rates.

It is worth mentioning that using multidimensional histograms is not the most efficient representation of a density function. The representation by a parameterized distribution for example would be more efficient since only a certain and typically small number of parameters needs to be stored. The dilemma is due to the tradeoff between representational efficiency and ability to discriminate. A basic goal of the article is to show that the representation of objects by the probability density function of their local image measurements contains enough discriminant information for the recognition of a variety of objects. Therefore we do not want to compromise on the ability to discriminate and have chosen multidimensional histograms for the estimation and representation of the probability function. Furthermore, multidimensional histograms provide us with a reliable estimate of the probability density function without being computational expensive. They also allow us to define simple and fast algorithms for recognition as histogram matching (section 5) and probabilistic object recognition algorithms (sections 6 and 8).

5 Histogram matching for recognition

Using a probability density function as an object representation allows the use of divergence functions from information theory and statistics [Basseville, 1996] directly for object recognition. Among these are e.g. the KL-divergence and the χ^2-divergence. We have experimentally compared such divergences to several histogram matching functions used in the computer vision literature [Schiele, 1997].

Let's assume the histogram of a test image is signified by $Q = \cup_{\mathbf{i}} q_{\mathbf{i}}$. Let $V = \cup_{\mathbf{i}} v_{\mathbf{i}}$ be a histogram from the object database. $\mathbf{i}$ is the L-dimensional index vector of a histogram, where L is the number of dimensions of a measurement vector m_k and therefore the number of dimensions of the histogram. $v_{\mathbf{i}}$ (respectively $q_{\mathbf{i}}$) corresponds to the value of a particular cell of histogram V (respectively Q).

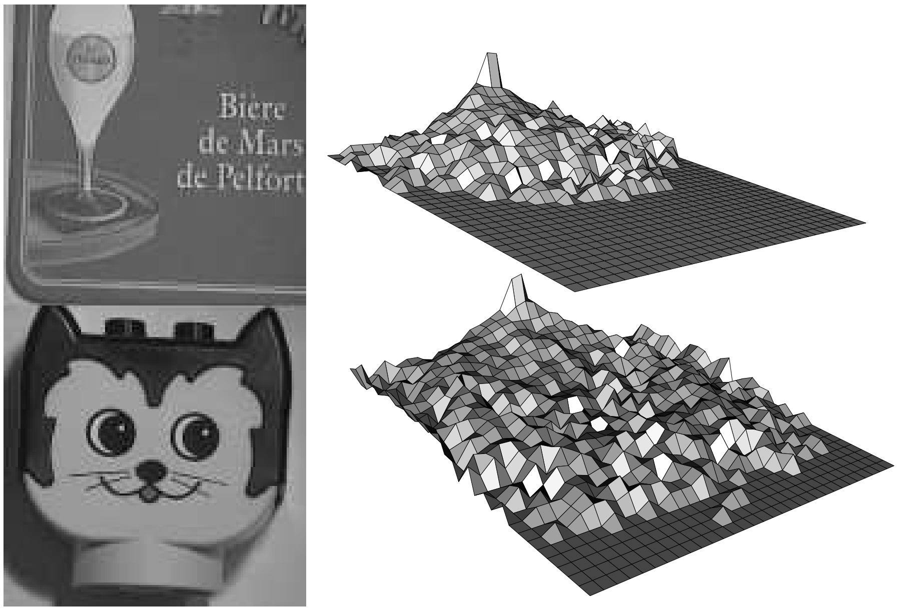

Figure 2: Two-dimensional histograms of two objects corresponding to a particular viewpoint, image plane rotation and scale. The image measurement is given by the Magnitude of the first derivative and the Laplace operator. The resolution of each histogram axis is 32.

The *intersection*-measurement [Swain and Ballard, 1991] has been introduced for the comparison of color histograms. The intersection of two histograms V and Q is defined by:

$$\cap(Q, V) = \sum_{\mathbf{i}} \min(q_{\mathbf{i}}, v_{\mathbf{i}}) \tag{21}$$

The intuitive motivation for this measurement is the calculation of the common part (the intersection) of two histograms V and Q. The main advantage of this measurement is that background pixels are neglected explicitly, which may occur in the test histogram Q but do not occur in the database histogram V.

In their original work Swain and Ballard reported the need for a sparse color distribution in the histogram in order to distinguish different objects. Our experiments have verified this result. A sparse distribution can be achieved by using high dimensional histograms. In this case the tradeoff between the ability to discriminate objects and stability with respect to perturbations becomes an important issue [Califano and Mohan, 1993]. A second inconvenience of the intersection is that all histogram cells are treated equally and should therefore be equally probable. This is approximately true for color histograms but cannot be assumed for the more general case of multidimensional receptive field histograms.

The χ^2-divergence is among the most prominent divergences used in statistics [Basseville, 1996] to assess the "dissimilarity" between two probability density functions. Two different ways of calculation of the χ^2–divergence may be considered [Press et al., 1992]. The first – $\chi^2_v(Q, V)$ – assumes exact knowledge of the model histogram V:

$$\chi^2_v(Q, V) = \sum_{\mathbf{i}} \frac{(q_{\mathbf{i}} - v_{\mathbf{i}})^2}{v_{\mathbf{i}}} \tag{22}$$

The second calculation – $\chi^2_{qv}(Q, V)$ – compares two observed histograms (neither is theoretically derived). This second χ^2–divergence is more appropriate in our context, since we do not assume exact knowledge of the model histogram V. $\chi^2_{qv}(Q, V)$ is defined by:

$$\chi^2_{qv}(Q, V) = \sum_{\mathbf{i}} \frac{(q_{\mathbf{i}} - v_{\mathbf{i}})^2}{q_{\mathbf{i}} + v_{\mathbf{i}}} \tag{23}$$

As we concluded from experiments [Schiele, 1997], these two χ^2–divergence provide better recognition results for most cases than the intersection measurement with respect to image distortions due to appearance changes, additive Gaussian noise and blur. Even though quadratic distances [Hafner et al., 1995] were typically outperformed by *intersection* and χ^2, the Mahalanobis distance – as a special case of quadratic distances – sometimes obtains comparable results [Schiele, 1997].

Figure 3: 25 of the 103 database objects use in the experiments.

Object recognition by means of histogram matching has been shown to be robust to appearance changes such as viewpoint changes, scale changes and image noise [Schiele, 1997]. This robustness is due to the fact that the proposed representation uses the *entire* appearance of the object rather than a small number of interest points. The appearance of objects is represented robustly by means of statistics of local neighborhood operators. As we will see in experiments (section 7) histogram matching also achieves a certain robustness to partial occlusion. However, histogram matching relies on some sort of pre-segmentation of the object. The next section proposes a probabilistic object recognition algorithm which calculates object hypotheses based on small image regions. This algorithm can be used successfully without using any pre-segmentation step.

6 Probabilistic recognition without correspondence

This section develops a probabilistic recognition technique which is based on single, arbitrarily chosen measurement vectors in the image. From such single measurement vectors the probability of the presence of each database object is calculated. The most noteworthy property of the algorithm is that the technique does not rely on the calculation of the correspondence between the test-image and the object database. In the following section, recognition results are given as a function of the visible object portion in order to show the robustness of the proposed probabilistic object recognition algorithm with respect to partial occlusion.

In the context of probabilistic object recognition we are interested in the calculation of the probability of an object o_n given a local image region R: $p(o_n|R)$. In our context, the most local region consists of a single local measurement vector m_k. This probability $p(o_n|m_k)$ can be calculated by the Bayes rule:

$$p(o_n|m_k) = \frac{p(m_k|o_n)p(o_n)}{p(m_k)} = \frac{p(m_k|o_n)p(o_n)}{\sum_i p(m_k|o_i)p(o_i)} \quad (24)$$

with

- $p(o_n)$ the *a priori* probability of object o_n,
- $p(m_k)$ the *a priori* probability of measurement vector m_k (= filter output combination),
- $p(m_k|o_n)$ the probability density function of object o_n. This density function can be estimated by the multidimensional receptive field histogram of an object o_n normalized by its size.

Typically, one single measurement vector will not be sufficient for the recognition of objects. Using two local measurement vectors m_k and m_j from the same object o_n we can calculate the probability of object o_n by:

$$p(o_n|m_k \wedge m_j) \quad = \quad \frac{p(m_k \wedge m_j|o_n)p(o_n)}{\sum_i p(m_k \wedge m_j|o_i)p(o_i)} \quad (25)$$

Under the assumption of *independence* of m_k and m_j we obtain:

$$p(o_n|m_k \wedge m_j) \quad = \quad \frac{p(m_k|o_n)p(m_j|o_n)p(o_n)}{\sum_i p(m_k|o_i)p(m_j|o_i)p(o_i)} (26)$$

Having K *independent* local measurement vectors $m_1, m_2, \ldots, m_K$ we can calculate the probability of each object o_n by:

$$p(o_n|\bigwedge_k m_k) \quad = \quad \frac{p(\bigwedge_k m_k|o_n)p(o_n)}{\sum_i p(\bigwedge_k m_k|o_i)p(o_i)} \quad (27)$$

$$= \quad \frac{\prod_k p(m_k|o_n)p(o_n)}{\sum_i \prod_k p(m_k|o_i)p(o_i)} \quad (28)$$

In our context the local measurement vectors m_k correspond to multidimensional receptive field vectors (for example two-dimensional vectors of the first Gaussian derivatives in the x– and y–directions). Therefore, K local measurement vectors m_k correspond to K receptive field vectors typically chosen from the same region of the image. It is worth mentioning that equation 28 assumes that all K measurement vectors come from the same object. This corresponds to an inherent consistency test which, as we will discuss later, is very powerful. However, regions with multiple objects may act as distractors to the algorithm. Experiments will show that already a small number of measurement vectors and therefore a small visible portion of an object provide reliable object hypotheses. More specifically, a visible object portion of 10% - 20% is generally enough in order to obtain good object hypotheses. That implies that the number of image regions containing a single object nearly always outnumbers the image regions containing multiple objects. The algorithm of section 8 makes use of this fact for the recognition of multiple objects in cluttered scenes where no pre-segmentation of the objects is assumed or used.

The a priori probabilities $p(o_n)$ of occurrence for each object o_n cannot be determined from the multidimensional receptive field histograms. These a priori probabilities depend upon the context and the given environment. Typically, they are constant for a certain context and/or environment. In the experiment of section 7 we assume that all objects are equally probable and do have a priori probabilities $p(o_n) = \frac{1}{N}$, with N the number of objects. Under this assumption equation 28 simplifies to:

$$p(o_n|\bigwedge_k m_k) \quad = \quad \frac{\prod_k p(m_k|o_n)}{\sum_i \prod_k p(m_k|o_i)} \quad (29)$$

As mentioned above, the probability density function $p(m_k|o_n)$ for an object o_n is directly given by its normalized multidimensional receptive field histogram. Therefore equation 29 shows a calculation of the probability for each object o_n entirely based on the multidimensional receptive field histograms of N objects.

It is important to note that the locations of the measurement vectors can be chosen arbitrarily. This is due to the fact that the position (t_x and t_y) of the measurement vectors are not represented in the object model (see section 4). Consequently the technique is fast (only a certain number of local receptive field vectors have to be calculated) and robust to partial occlusion (the approach is strictly local). Furthermore, the technique works *without* calculation of the correspondence between the object database and the test image.

7 Experimental results

The section describes an experiment using a database of 2130 images of 103 different objects. Figure 3 shows some of the database objects. We have taken 690 different images of 83 objects where each of the images correspond to a different scale and different rotation of the object in front of the camera. See figure 4 for examples of different scales. The remaining 1440 images come from the Columbia image database which contains 72 viewpoints of 20 different objects [Murase and Nayar, 1995].

In this experiment we use six–dimensional histograms of the filter combination *Dx-Dy* (first Gaussian derivative in x and y directions) at three different scales with $\sigma_1 = \sigma$, $\sigma_2 = 2\sigma$ and $\sigma_3 = 4\sigma$. The resolution per histogram axis is 24 (see for details of the estimation section 7.1).

The training set for the 83 objects contains one image for each object. For each of these images we calculate a set of histograms corresponding to different scales and image plane rotations of the object. By making use of the steerability of the Gaussian derivatives we calculate histograms which correspond to different image plane rotations from a single image per object. Similarly we use the equivariance property of the Gaussian derivatives to scale changes to obtain histograms which correspond to different scales of an object. We calculate histograms of 6 different scales covering the approximate scale factor of 2.2 for the test images. For each of these scales we also calculate histograms for 18 different image rotations covering 360^o degrees image plane rotation[4]. The overall number of histograms for the 83 objects is therefore $83 \times 18 \times 6 = 8964$ histograms. These histograms are stored in the histogram database.

The Columbia image database has been created by [Murase and Nayar, 1995] and used by several researchers including [Rao and Ballard, 1995] [Schmid and Mohr, 1997]. As mentioned above, the database contains 72 viewpoints for each of the 20 objects. The viewpoints are 5^o apart. Typically, every other viewpoint is taken as training image and the remaining images are taken as test set. The training set as well as the test set contain 720 images. For each training image we calculate one histogram corresponding to the particular rotation and scale of the object. This adds 720 histograms to the histogram database. The total number of histograms in the database is therefore $8964 + 720 = 9684$.

The test set contains the remaining images of the 83 objects which is $690 - 83 = 607$ and 720 images of the Columbia image database. The total number of test images is therefore 1327. The entire test set is independent of the training images.

In order to recognize the objects in the test images we calculate one six–dimensional histogram with

[4]More specifically for each of the 83 images of 83 objects we calculate histograms which correspond to 18 different rotations namely for the angles of $\alpha = 0°, 20°, 40°, \ldots, 340°$. For each of these rotations we calculate histograms which correspond to 6 different scales namely $\sigma = 1.48, 1.7, 2.0, 2.26, 2.62$ and 3.0. This range of the σ's is motivated by the maximum scale factor of 2.2 which we used.

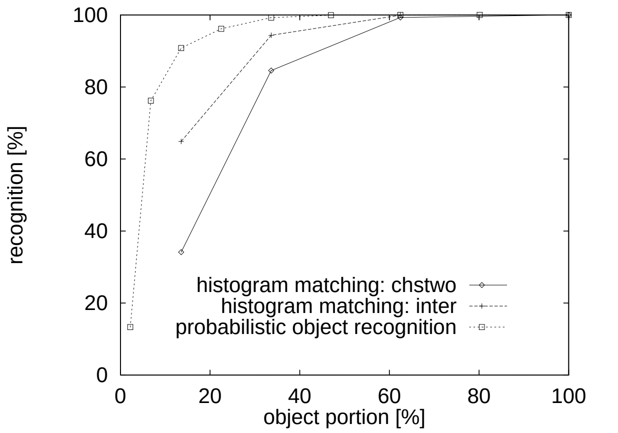

Figure 5: Experimental results for 1327 test images of 103 objects. Comparison of probabilistic object recognition and recognition by histogram matching: χ^2_{qv} (chstwo) and $\cap$ (inter)

$\sigma = 2.0$ per test image. The support of these histograms is varied (from about 20% to 100% visibility of the objects) in order to test the robustness of the approach to partial occlusion. Since the objects are centered in the image we have calculated the histograms of a centered support region. This corresponds to the ideal case that the object position is approximately known. Figure 5 shows the recognition results obtained by two different histogram comparison measurements: χ^2_{qv} and $\cap$ (see section 5). The recognition result is shown as a function of the visible portion of the objects.

Figure 5 shows a 100% recognition provided by both comparison measurements using the entire object as support for the histogram calculation. By using only 62% of the object the intersection measurement still provides 100% recognition. In this case χ^2_{qv} obtains 99.3% recognition. In the case of 33% visibility of the object, $\cap$ provides a recognition of 94% and χ^2_{qv} obtains 84% recognition. The experiment emphasizes in particular the expected robustness of the intersection measurement $\cap$ with respect to partial occlusion.

This initial experiment shows the applicability of histogram matching for object identification in the presence of scale changes, image plane rotation, viewpoint changes and partial occlusion. In particular, these results emphasize that multidimensional histograms represent the appearances of objects reliably enough in order to discriminate 100 objects.

In order to apply the probabilistic object recognition algorithm (equation 29), K independent measurement vectors m_k have to be chosen from a test image. As mentioned above, two assumptions underlying equa-

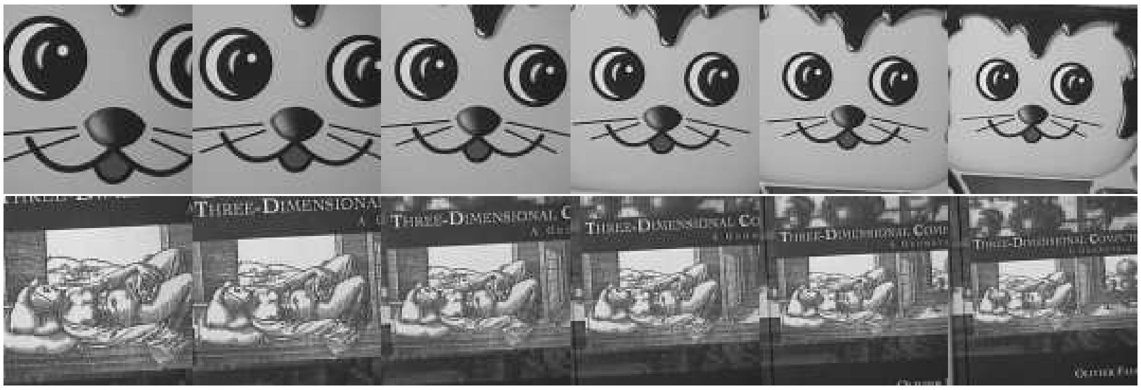

Figure 4: Six different scale-images for 2 objects which are part of the test-set.

tion 29 have to be considered: firstly, all measurement vectors are assumed to correspond to the same object and secondly, the K measurement vectors m_k are assumed to be independent. The second assumption, the independence of the measurement vectors, is fulfilled by using a fixed distance of 4 pixels between each measurement vector corresponding to $2\sigma_1$, which is sufficient to assume independence from a signal processing point of view. The first assumption is satisfied here by using test images containing only one object and choosing the measurement vectors from a central region of the test images. The reported results therefore correspond to the ideal case that all K measurements come from the same object. In general there is no trivial way in which to satisfy the first assumption. Nevertheless, the experimental results reported below indicate that a small object portion is sufficient for a good object hypothesis. This property of the algorithm is used in section 8 in order to extend the algorithm for the recognition of multiple objects in cluttered scenes and without segmentation.

Figure 5 and table 1 summarize the recognition results of the probabilistic object recognition algorithm. A visible object portion of approximately 62% is sufficient for the recognition of all 1327 test images (the same result as for histogram matching). With 33.6% visibility the recognition rate is above 99% (10 errors in total). Using 13.5% of the object the recognition rate is still above 90%. The recognition rate is 76% with only 6.8% visibility of the object. This can be explained by the fact that each single vector contains discriminant information. This is stressed also by a recognition of approximately 13% with only a single measurement vector.

Since we use the same six–dimensional feature vectors as for the recognition by histogram matching as for the probabilistic recognition algorithm, we can directly compare the results of both algorithms in figure 5. As we can see the robustness to partial occlusion is significantly increased by applying the probabilistic object recognition scheme.

We can conclude that the proposed probabilistic object recognition approach is capable of discriminating 103 objects in the presence of significant scale changes, image plane rotation and viewpoint changes. Furthermore, the approach is robust with respect to partial occlusion since a small portion of the object is sufficient in order to obtain a good object hypothesis. As mentioned earlier, the recognition results have been obtained without any correspondence calculation between the test images and the database.

7.1 Implementation details

For the experiments described in this section the resolution of each histogram axis has been 24. Therefore the theoretical number of cells for a six-dimensional histogram is in the order of 10^8 cells. Due to the dependencies between the different dimensions of the histogram axes and due to the fact that not all theoretical possible pixel-values are observed in real images, the number of non-zero histogram cells (for all 9684 histograms) is in the order of 10^6. This is number is still too large to be estimated from a typical 512×512 image which contains about 2×10^5 pixels. However, by using an appropriate bias for the histograms (in our case a uniform prior) we can effectively decrease the number of cells to be estimated below the order of 10^5. This prior is important to ensure a reliable estimate of the multidimensional histograms. In reality, however, the exact amount of the prior only has a secondary effect [Schiele, 1997] on the performance of the algorithm.

The test-set contains also images of different scales (see figure 4 for two examples). In order to calculate histograms of filter responses at arbitrary scales we apply

radius [σ_1]	1	5	10	15	20	25	30	35	40
object portion [%]	2.2	6.8	13.5	22.5	33.6	47.0	62.5	80.1	100.0
number of image measurements	1	25	100	225	400	625	900	1225	1600
recognition [%]	13.3	76.2	90.8	96.2	99.3	99.9	100	100	100
errors for the 83 objects	577	274	122	51	10	1	0	0	0
errors for Columbia database	573	42	0	0	0	0	0	0	0

Table 1: Experimental results with 103 objects

two principles: firstly we use the equivariance property of Gaussian derivatives to scale and secondly we adapt the radius of the support region of a histogram as a function of scale. The equivariance property is described in section 3.2. In order to calculate the histogram of vectors of Gaussian derivatives of a set of image positions, we need to adapt the image positions of the vectors. This can be done for example by the adaptation of the distances between image positions, which would include interpolation between pixels. Due to the computational cost of interpolation, we prefer to leave the pixel distances constant and to adapt the support region for the calculation of the histogram. The radius of the support region needs to be multiplied by the scale. This adaptation of the support region is computationally inexpensive but compromises the precision in particular for small scales. Therefore histograms corresponding to different scales of an image are calculated on different support regions and contain different numbers of entries. In order to make such histograms comparable the overall sum of the histogram entries needs to be normalized.

For histograms steered to different rotations, the support region should be circular. In contrast to a circular support region, a square region – using the same radius as half side-length of the square – contains about 20% more measurement vectors which is advantageous for the small radii used here (see above). Fortunately, imprecision due to square support regions are introduced only for the borders of the objects. In this experiment we use square, small and centered support regions. The size of the support regions is limited by the image sizes. Since we calculate histograms at different scales of objects the maximal possible radius of the support region is $40\sigma_1$. This radius corresponds to a radius of 59 pixels (for $\sigma_1 = 1.48$) and 120 pixels (for $\sigma_1 = 3.0$). Therefore the support region of the histograms differs up to a factor of $4 \approx \frac{120^2}{59^2}$. The centering of the support region can be seen as a figure–ground segmentation for learning an object model.

8 Multiple object recognition in cluttered scenes

In the previous section we applied the probabilistic algorithm for the recognition of *single* objects in the presence of partial occlusion. As mentioned earlier hashtable based recognition systems are very suited for the recognition of *multiple* objects in cluttered scenes. Motivated by the results of the preceding section we can define an algorithm for the recognition of multiple objects which employs local image regions or local appearances of objects for probabilistic voting for objects. Since this resembles to use local appearance as index of a hashtable we will call this algorithm *local appearance hashing*.

The upper part of figure 6 shows the standard hashtable approach: for each feature vector m_i the approach votes for a certain subset of objects denoted by $vote(o_n|m_i)$: this vote is one if object o_n could correspond to the feature vector m_i and zero otherwise. The votes for an object are summed over the entire image: $votes(o_n|Image) = \sum_i vote(o_n|m_i)$.

This hashtable algorithm typically produces a high number of false positives. In order to overcome this problem we can use pairs or triplets of feature vectors and their geometric arrangement to increase the discriminant power of the approach [Schmid and Mohr, 1997]. Another possibility is to increase the dimensionality of the feature vector [Rao and Ballard, 1997] resulting in an enlarged support region for the feature vector. These approaches pursue interesting directions by coding additional geometrical or consistency constraints. Eventually, we will integrate these ideas into our framework. However, the main disadvantage of these approaches is that the additional constraints have to be coded into the hashtable prior to recognition. Therefore, motivated by the results of the previous section, we will make use of the discriminant power of the statistically distribution of the feature vectors.

The probabilistic algorithm defined in section 6 is structurally similar to a hashtable based algorithm (see lower part of figure 6). In the probabilistic algorithm, we calculate for each feature vector m_i the probabili-

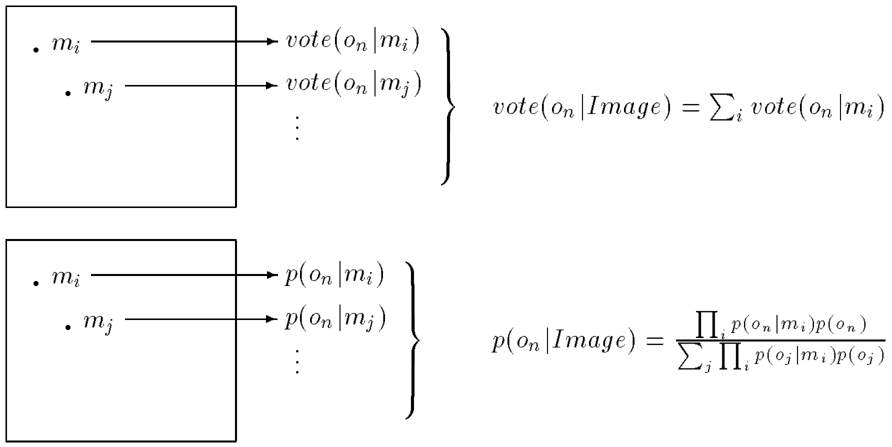

Figure 6: Comparison of (above) hashtable based recognition and (below) the probabilistic recognition of section 6.

ties $p(o_n|m_i)$. The evidence for an object in the image $p(o_n|Image) = p(o_n|\bigwedge_i m_i)$ is accumulated using equation 28 or 29 respectively. In any case, since all feature vectors m_i are assumed to come from the same object this is equivalent to an inherent consistency test using the distribution of the feature vectors. As the results of the previous section show this is a powerful consistency constraint. However, this algorithm is not suited to recognize multiple objects in cluttered scenes.

The results of the previous section indicate that a relatively small region is sufficient in order to obtain a good object hypothesis. By making use of this property of the algorithm and combining it with a hashtable we obtain a hybrid algorithm which combines the advantages of both. Figure 7 shows this hybrid algorithm. Instead of accumulating the evidence of each object over the entire image we apply the probabilistic algorithm only for a local image region R_k and calculate the corresponding probabilities $p(o_n|R_k) = p(o_n|\bigwedge_{k_i} m_{k_i})$ (where the m_{k_i} correspond to the feature vectors inside region R_k). Calculating these probabilities for a set of image regions R_k we can accumulate the evidence for each object o_n by $votes(o_n|Image) = \sum_k p(o_n|R_k)$. This last step corresponds to using image regions R_k as "feature vectors" in a hashtable. Since these local image regions correspond to local appearances of the objects we call this approach *local appearance hashing*.

We like to point out an interesting property of the proposed *local appearance hashing* approach. Since the regions R_k can be chosen arbitrarily and dynamically during runtime, the algorithm is extremely flexible. In particular, the size and form of the local image regions R_k can be changed dynamically without recalculating the representation of the objects. Since these image regions correspond to the "feature vectors" we can actually change these feature vectors dynamically, depending e.g. what we know about the scene. For any chosen image region R_k the algorithm implicitly uses the consistency constraint imposed by the distribution over the feature vectors for each object.

8.1 Recognition experiment

In order to illustrate the proposed local appearance hashing approach we describe an experiment on a database of 50 objects. For each of the 50 objects we compute six–dimensional histograms *Mag-Lap-24* (Magnitude of first derivative and Laplacian operator, resolution of 24 cells per histogram axes) at three different scales, namely $\sigma_1 = 2.0, \sigma_2 = 4.0$ and $\sigma_3 = 8.0$[5]. For illustration purposes, the image regions R_k have been fixed to a squared region of 64^2 pixels. We have chosen $6 \times 6 = 36$ such regions overlapping the neighboring regions by 50%. For each of the 36 regions we apply the probabilistic object recognition algorithm and add the computed probabilities into an accumulator array of the objects. Objects, which cover several image regions R_k therefore accumulate probabilities of several image regions. The more image regions are covered by an object the higher the score becomes. Ultimately, the objects with the highest "scores" in the accumulator are listed in decreasing order (see figure 8).

We have taken a set of 50 test images each containing 3 of the 50 objects in order to test the performance of the algorithm. The left column of figure 8 shows 4 of these test images. For each of these test images the objects with the highest "scores" are displayed. The first three

[5]The remarks made in section 7.1 about the estimation of the multidimensional histograms also applied here

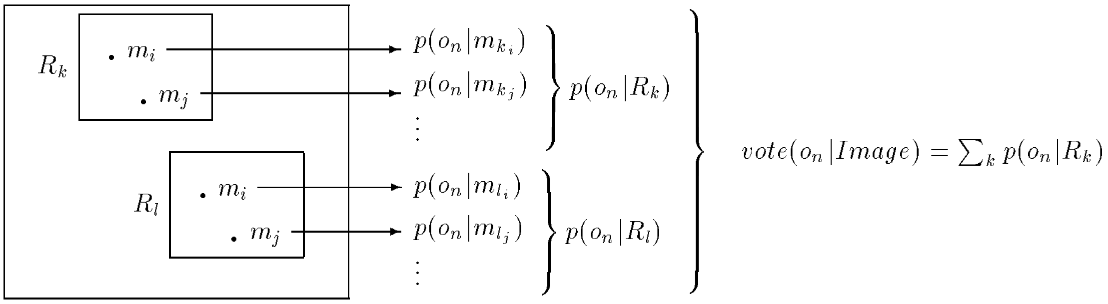

Figure 7: *Local appearance hashing*: combining the probabilistic recognition algorithm of section 6 with a hashtable in order to recognize multiple objects in cluttered scenes.

# of matches	1	2	3	4	5	6 ... 14
1 object correct	47	50	50	50	50	50 ... 50
2 objects correct		40	49	50	50	50 ... 50
3 objects correct			27	45	48	49 ... 50
overall			126	145	148	149 ... 150

Table 2: Recognition results for 50 test images containing 3 objects

matches for each of the first three test images contain all three objects which are contained in the image. For the fourth test image the first two and the fourth match are correct. However, even though the third match is not contained in the test image it corresponds to a similar object as the first match. This illustrates the property of the algorithm that it tents to match visually similar objects. Table 2 summarizes the results for the 50 test images. As we can see many of the objects (126 of 150) are contained within the first three matches. By including four matches 145 of the possible 150 objects are recognized. Since the results have been obtained only for a small set of test images it is unreasonable to generalize them. However, the results clearly indicate the possibility to recognize multiple objects in cluttered scenes using the proposed local appearance hashing approach.

The first row of figure 9 shows another set of interesting test images. Each of these test images contains one of the 50 objects of the database. The rest of the images is covered by objects which are *not* part of the database and therefore are not represented. These types of images are considered difficult in particular for probabilistic object recognition algorithms since they typically rely on the assumption that they have a complete model of the world. Even though this assumption is shared by our probabilistic algorithm the local appearance hashing approach recognizes the correct object three times as best match (test images A, B and D) and once as third best match (test image C). This ability to recognize objects in the presence of not represented objects is mainly due to the consistency constraint which is implicitly imposed by the use of the distribution of the feature vectors.

9 Conclusions

For nearly forty years, the field of computer vision has struggled with the techniques for recognizing complex objects by searching correspondences between object models, and local structure in images. Recognition using correspondence between models and images has proved both computationally expensive and sensitive to image noise. In almost every case, model based recognition techniques required a small pre–selected list of candidate objects in order to be tractable. The general assumption has been that the candidates would be provided by context.

Recognition using joint statistics of local properties provides an alternative to standard recognition algorithm. This approach provides a framework in which it is possible to design techniques to determine the objects in a scene independent of viewing position. These techniques have computational complexities which are linear with the number of pixels and the number of objects, and thus can be implemented to operate in real time. Indeed, we have implemented an example of such a system which operates at 10Hz on a standard workstation with a data base of 103 objects.

This framework can be used with a large variety of local properties. However, linear filters based on the Gaussian function are particularly well suited as they permit the definition of local property measurements which are robust to changes in scale and orientation. In particular, our experiments have shown excellent results with local properties measured using Gaussian derivatives at different scales and Gabor filters at different

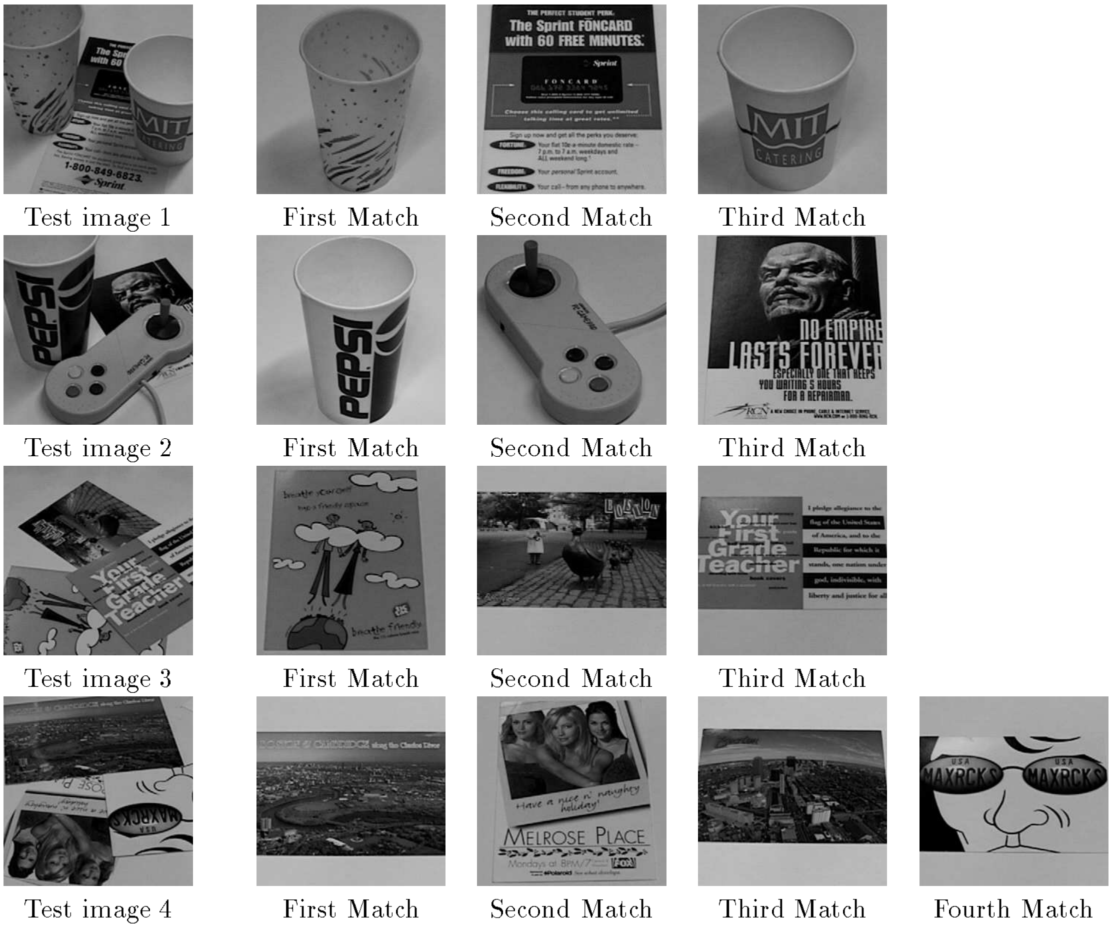

Figure 8: Four of the 50 test images containing multiple objects.

scales. The steerability property of such operators is especially useful in providing an efficient means to obtain image plane rotation invariant recognition.

Histograms of local property vectors provide a robust and simple means to answer the question: What is the probability that the pixels in a region of an image contain a projection from an object? A probabilistic approach has proven particularly reliable formulation for this process. Probabilistic recognition from joint statistics of local properties is robust to occlusions and cluttered scenes.

These results demonstrate that the appearance of an object is the composition of the appearance of its parts. Thus object appearance is best captured as a composition of local appearances, as measured by a vector of local operators such as Gabor filters or Gaussian derivatives. The joint statistics of local appearance measures provide a powerful basis for object indexing and recognition. This approach is complementary to a structural description of local appearance.

Acknowledgments

We would like thank the anonymous reviewers for their constructive comments and suggestions which helped improving the quality of the paper.

10 References

[Ballard and Rao, 1994] Ballard, D. and Rao, R. (1994). Seeing behind occlusions. In *ECCV'94 Third European Conference on Computer Vision, Volume I*, pages 274–285.

[Ballard and Wixson, 1993] Ballard, D. and Wixson, L. (1993). Object recognition using steerable filters at multiple scales. In *IEEE Workshop on Qualitative Vision*, pages 2–10.

[Basseville, 1996] Basseville, M. (1996). Information: entropies, divergences et moyennes. Technical Report 1020, IRISA. in French.

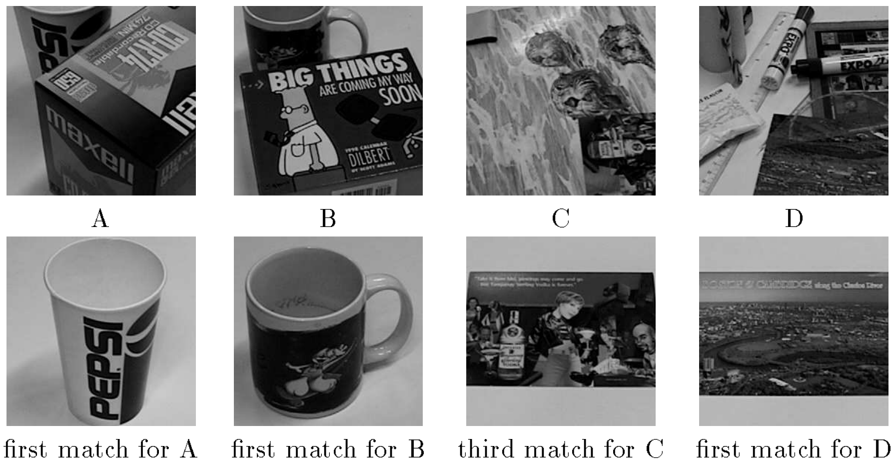

Figure 9: Four test images with objects of the database and objects which are *not* represented in the database.

[Belongie et al., 1998] Belongie, S., Carson, C., Greenspan, H., and Malik, J. (1998). Color- and texture-based image segmentation using the expectation-maximization algorithm and its application to content-based image retrieval. In *ICCV'98 Sixth International Conference on Computer Vision*, pages 675–682.

[Burkhardt and Zisserman, 1992] Burkhardt, H. and Zisserman, A., editors (1992). *Invariants for Recognition.* ESPRIT–Basic-Research-Workshop, ECCV'92.

[Burns et al., 1990] Burns, J., Weiss, R., and Riseman, E. (1990). View variation of point set and line segment features. In *Proceedings DARPA Image Understanding Workshop*, pages 650–659.

[Califano and Mohan, 1993] Califano, A. and Mohan, R. (1993). Systematic design of indexing strategies for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 709–710.

[Clemens and Jacobs, 1991] Clemens, D. and Jacobs, D. (1991). Space and time bounds of indexing 3-d models from 2-d images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1007–1017.

[Daugman, 1993] Daugman, J. (1993). High confidence visual recognition of persons by test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1148–1161.

[Deriche, 1987] Deriche, R. (1987). Using canny's criteria to derive a recursively implemented optimal edge detector. *International Journal of Computer Vision*, 1(2):167–187. see also [Deriche, 1993].

[Deriche, 1993] Deriche, R. (1993). Recursively implementing the gaussian and its derivatives. Technical Report 1893, INRIA–Sophia Antipolis.

[Ennesser and Medioni, 1995] Ennesser, F. and Medioni, G. (1995). Finding waldo, or focus of attention using local color information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):805–809.

[Finlayson et al., 1998] Finlayson, G., Schiele, B., and Crowley, J. (1998). Comprehensive colour image normalization. In *ECCV'98 Fifth European Conference on Computer Vision, Volume I*, pages 475–490.

[Flickner et al., 1995] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Juang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. (1995). Query by image and video content: The QBIC system. *IEEE Computer*, pages 23–32.

[Freeman and Adelson, 1991] Freeman, W. and Adelson, E. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906.

[Fukunaga, 1990] Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition.* Computer Science and Scientific Computing. Academic Press, New York, 2nd edition.

[Funt and Finlayson, 1995] Funt, B. and Finlayson, G. (1995). Color constant color indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):522–529.

[Gabor, 1946] Gabor, D. (1946). Theory of communication. *Proc. Inst. Elec. Eng.*, 93(26):429–441.

[Grimson et al., 1994] Grimson, W., Huttenlocher, D., and Jacobs, D. (1994). A study of affine matching with bounded sensor error. *International Journal of Computer Vision*, 13(1):7–32.

[Grimson and Huttenlocher, editors, 1991] Grimson, W. and Huttenlocher, editors, D. (1991). Special issue on interpretation of 3-d scenes – part i. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10).

[Grimson and Huttenlocher, editors, 1992] Grimson, W. and Huttenlocher, editors, D. (1992). Special issue on interpretation of 3-d scenes – part ii. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2).

[Hafner et al., 1995] Hafner, J., Sawhney, H., Equitz, W., Flickner, M., and Niblack, W. (1995). Efficient color histogram indexing for quadratic form distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7):729–736.

[Haralick, 1979] Haralick, R. (1979). Statistical and structural approaches to texture. *Proceedings of IEEE*, 67(5):786–804.

[Healey and Slater, 1994] Healey, G. and Slater, D. (1994). Using illumination invariant color histogram descriptors for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 355–360.

[Hornegger and Niemann, 1995] Hornegger, J. and Niemann, H. (1995). Statistical learning, localization and identification of objects. In *ICCV'95 Fifth International Conference on Computer Vision*, pages 914–919.

[Intrator and Gold, 1993] Intrator, N. and Gold, J. (1993). Three–dimensional object recognition using an usupervised bcm network: The usefulness of distinguishing features. *Neural Computation*, 5:61–74.

[Jones and Malik, 1992] Jones, D. and Malik, J. (1992). A computational framework for determining stereo correspondence from a set of linear spatial filters. In *ECCV'92 Second European Conference on Computer Vision*, pages 395–410.

[Koenderink and Doorn, 1987] Koenderink, J. and Doorn, A. (1987). Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375.

[Lamdan et al., 1988] Lamdan, Y., Schwartz, J., and Wolfson, H. (1988). Object recognition by affine invariant matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 335–344.

[Lamdan and Wolfson, 1988] Lamdan, Y. and Wolfson, H. (1988). Geometric hashing: A general and efficient model based recognition scheme. In *ICCV'88 Second International Conference on Computer Vision*, pages 238–249.

[Malik and Perona, 1989] Malik, J. and Perona, P. (1989). A computational model of texture segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–332.

[Mao and Jain, 1992] Mao, J. and Jain, A. (1992). Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern Recognition*, 25(2):173–188.

[Matas et al., 1995] Matas, J., Marik, R., and Kittler, J. (1995). On representation and matching of multi–colored objects. In *ICCV'95 Fifth International Conference on Computer Vision*, pages 726–732.

[Mel, 1997] Mel, B. (1997). Seemore: Combing color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation*, 9:777–804.

[Moghaddam and Pentland, 1995] Moghaddam, B. and Pentland, A. (1995). Maximum likelihood detection of faces and hands. In *International Workshop on Automatic Face– and Gesture–Recognition*, pages 122–128.

[Mohr et al., 1997] Mohr, R., Picard, S., and Schmid, C. (1997). Bayesian decision versus voting for image retrieval. In *Procceedings pf the 7th International Conference on Computer Analysis of Images and Patterns*, pages 376–383.

[Mundy and Zisserman, 1992] Mundy, J. L. and Zisserman, A., editors (1992). *Geometric Invariance in Computer Vision*. MIT Press.

[Mundy et al., 1993] Mundy, J. L., Zisserman, A., and Forsyth, D., editors (1993). *Application of Invariance in Computer Vision*. Number 825 in Lecture Notes in Computer Science. Springer Verlag.

[Murase and Nayar, 1995] Murase, H. and Nayar, S. (1995). Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, 14:5–24.

[Nagao, 1995] Nagao, K. (1995). Recognizing 3d objects using photometric invariants. In *ICCV'95 Fifth International Conference on Computer Vision*, pages 480–487.

[Object representation, 1996] Object representation (1996). International workshop on object representation for computer vision. Cambridge, England.

[Ohba and Ikeuchi, 1996] Ohba, K. and Ikeuchi, K. (1996). Recognition of the multi specularity objects for bin–picking task. In *IROS'96 Intelligent Robots and Systems*, pages 1440–1447, Osaka, Japan.

[Perona, 1995] Perona, P. (1995). Deformable kernels in early vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):488–499.

[Popat and Picard, 1994] Popat, K. and Picard, R. (1994). Cluster–based probability model applied to image restoration and compression. In *IEEE Conference on Acoustics, Speech and Signal Processing*.

[Pope, 1995] Pope, A. (1995). *Learning to Recognize Objects in Images: Acquiring and Using Probabilistic Models of Appearance*. PhD thesis, Department of Computer Science, University of British Columbia.

[Pope and Lowe, 1996] Pope, A. and Lowe, D. (1996). Learning appearance models for object recognition. In *International Workshop on Object Representation for Computer Vision*, Cambridge, England.

[Press et al., 1992] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge University Press, 2nd edition.

[Rao and Ballard, 1995] Rao, R. and Ballard, D. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence*, 78:461–505.

[Rao and Ballard, 1997] Rao, R. and Ballard, D. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, 9(4):721–763.

[Rigoutsos and Hummel, 1993] Rigoutsos, I. and Hummel, R. (1993). Distributed bayesian object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 180–186.

[Schiele, 1997] Schiele, B. (1997). *Object Recognition using Multidimensional Receptive Field Histograms*. PhD thesis, I.N.P.Grenoble. English translation.

[Schmid and Mohr, 1997] Schmid, C. and Mohr, R. (1997). Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535.

[Schmid et al., 1998] Schmid, C., Mohr, R., and Bauckhage, C. (1998). Comparing and evaluating interest points. In *ICCV'98 Sixth International Conference on Computer Vision*.

[Sirovich and Kirby, 1987] Sirovich, L. and Kirby, M. (1987). Low–dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America*, 4(3):519–524.

[Slater and Healey, 1995] Slater, D. and Healey, G. (1995). Combining color and geometric information for the illumination invariant recognition of 3d objects. In *ICCV'95 Fifth International Conference on Computer Vision*, pages 563–568.

[Swain and Ballard, 1991] Swain, M. and Ballard, D. (1991). Color indexing. *International Journal of Computer Vision*, 7(1):11–32.

[Tsotsos, 1989] Tsotsos, J. (1989). The complexity of perceptual search tasks. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, pages 1571–1577.

[Turk and Pentland, 1991] Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86.

[Westelius, 1992] Westelius, C.-J. (1992). *Preattentive Gaze Control for Robot Vision*. PhD thesis, Department of Electrical Engineering, Linköping University.

[Wolfson, 1990] Wolfson, H. (1990). Model-based object recognition by geometric hashing. In *ECCV'90 First European Conference on Computer Vision*, pages 526–536.

[Young, 1986] Young, R. (1986). Simulation of human retinal function with the gaussian derivative model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 564–569.