THESE

présentée par

Bernt SCHIELE

pour obtenir le grade de DOCTEUR

de l'INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

(Arrêté ministériel du 30 mars 1992)

spécialité INFORMATIQUE

Object Recognition using Multidimensional Receptive Field Histograms

English translation

Date de soutenance : 16 juillet 1997

Composition du jury :

Président :	Roger Mohr
Rapporteurs :	Shimon Edelman
	Henri Maître
Examinateurs :	Hans Burkhardt
	James L. Crowley (Directeur de thèse)

Thèse préparée au sein du laboratoire GRAVIR - IMAG
I.N.P. de Grenoble, 46, av. Félix Viallet, 38031 Grenoble Cedex, FRANCE

Acknowledgments

I wish to express my sincere gratitude to all those people who have contributed in different ways to the work that is presented in this thesis.

First of all I would like to thank my advisor and co–worker, Prof. James L. Crowley, who provided me motivation, a stimulating research environment and who introduced me to the research community. I am also grateful to Roger Mohr, Shimon Edelman, Henri Maître and Hans Burkhardt for the interest in my work and for being members of the thesis jury.

For friendship and research collaboration, thanks to Stephen Jones, Guido Appenzeller and Frank Wallner. I am especially grateful to Steve who has patiently corrected the English version of the thesis.

Many thanks to many people for smoothing the rough edges of my French and in particular to Bénédicte, Beatrice, Serge, Jérôme, Christophe, Vincent, Augustin, Alexandre, Delphine, Claude et Vincent le deuxième.

Thanks to all the current and past members of the research group PRIMA, Alvaro, Bénédicte, Bruno, Claude, Claus, Christophe, Cordelia, Frank, Jérôme, Jim, Olivier, Philippe, Patrice, Patrick, Steve, Vincent and Vincent le deuxième, for the friendly, fun and international atmosphere. It has been a pleasure working in Grenoble. Special thanks to Philippe for his help and discussions in the first months.

No words are adequate to express the author's debt to his parents, his sister, his two brothers and their families for their love and support. The work is dedicated to the author's brother Ralf who has been always a source of encouragement.

Finally I would like to thank you, Silke, for your patience, love and support and for just being.

Bernt Schiele,
Grenoble, 21 July 1997

Abstract

During the last few years, there has been a growing interest in object recognition techniques directly based on images, each corresponding to a particular appearance of the object. Representations of objects, which use only information of images are called *appearance based* models. The interest in such representation schemes is due to their robustness, speed and success in recognizing objects.

The thesis proposes a framework for the statistical representation of the appearances of 3D objects. The representation consists of a probability density function over a set of robust local shape descriptors which can be extracted reliably from images. The object representation is therefore learned automatically from sample images. Multidimensional receptive field histograms are introduced for the approximation of the probability density function. A main result of the thesis is that such a representation scheme based on local object descriptors provides a reliable means for object representation and recognition.

Different recognition algorithms are proposed and experimentally evaluated. The first recognition algorithm by histogram matching can be seen as a generalization of the color indexing scheme of Swain and Ballard. The second recognition algorithm calculates probabilities for the presence of objects only based on multidimensional receptive field histograms. The most remarkable property of the algorithm is that it relies on neither correspondence nor figure ground segmentation. Experiments show that this algorithm is capable of recognizing 100 objects in cluttered scenes. The third recognition algorithm incorporates several viewpoints in an active recognition framework in order to solve ambiguities inherent in single view recognition schemes.

The thesis also proposes visual classes as a general framework for appearance based object classification. Classification has been proven difficult for arbitrary objects due to instabilities of invariant representations. The proposed concepts for extraction, representation and recognition of visual classes provide a general framework for object classification.

From an abstract point of view, the thesis aims to push the limits of the appearance based paradigm without using neither figure ground segmentation nor correspondence. The active object recognition method allows the consistent recognition of objects in 3D and therefore overcomes the limits of single view recognition. The appearance based classification framework based on the concept of visual classes will serve for future research.

Table of Contents

Preface

In the classical approach to image analysis approach [Mar 78, Kan 78, Ros 84, Nag 92] *image features* such as edges or *image regions* such as texture regions are extracted from the image. High level *feature groups* may be obtained by *grouping* these basic image features. This approach hypothesizes the identity and the pose of the object in the scene by calculating *feature correspondence* between the feature groups and the features of the object model.

The principal difficulty with this classical approach is that the process of determining feature correspondence has a complexity which is exponential with the number of extracted image features. Furthermore the extraction and grouping processes which produce image features are unstable, producing broken and spurious features which compound the complexity of correspondence.

In order to make the problem tractable, the number of extracted features must be reduced. This implies the use of salient – meaning discriminant – features. Because of the exponential complexity, only a relatively small number of image features can be used so that each image feature must be highly discriminant. Due to the tradeoff between robustness of the feature extraction and the discriminant power of features, the process of feature extraction tends to be unstable. Furthermore, the saliency of image features depends on the object classes employed making the techniques suitable only for particular object classes such as geometric objects.

The above limitations of the classical approach to image analysis require a paradigm shift in computer vision: the object's identity and the object's pose are estimated directly from measurements which can be calculated reliably from the image. The process of estimating the object's identity and the object's pose has a complexity which can be linear with the number of image measurements. This implies that a large number of image measurements may be used and therefore that robust image measurements can be chosen. In this context, the model of an object is given by a representation of image measurements which can be learned automatically from sample images. These techniques are called appearance based methods since each of the represented images corresponds to a particular appearance of the object.

The advantage of appearance based methods is that they can use robust image measurements and that they can avoid feature correspondence. From an abstract point of view, these techniques calculate *object correspondences* between the image and the object models. This calculation of object hypotheses might be used as a pre–step of the classical image analysis approach: the hypothesized object can serve as a priori knowledge in order to reduce the complexity of the processes of correspondence and grouping.

Different appearance based object recognition techniques have been proposed: examples include the alignment scheme of Huttenlocher and Ullman [Hut 87], which relies on point correspondence of a small number of salient features, the eigenpicture approach [Sir 87, Tur 91a, Mur 95], which assumes the detection or the segmentation of the object, and the aspect graph [Koe 79, Fau 92, Egg 93], which is so far only applied to geometric objects.

The color indexing approach of Swain and Ballard [Swa 91] uses directly the color distribution of objects for recognition. Their approach 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. This approach is an attractive method for object recognition, because of its simplicity, speed and robustness. However, its reliance on object color and, to a lesser degree, light source intensity make it inappropriate for many recognition problems.

The focus of our work has been to develop a technique similar to color indexing using local descriptions of an object's shape provided by a vector of linear neighborhood operators. The first part of the thesis is therefore concerned with the definition of a statistical object representation framework based on local neighborhood operators. The principal aim is to develop fast and robust recognition techniques using the defined statistical object representation. The applicability of the techniques is shown experimentally on different databases each containing up to 100 objects. In order to overcome the limitations of the classical image analysis approach the thesis examines recognition without reliance on pre–segmentation and feature correspondence.

The speed and the robustness of appearance based object recognition approaches comes with a price: appearance based approaches use directly image measurements for recognition. Images and therefore appearances are recognized rather than objects. Due to this fact, any appearance based approach has to be evaluated with respect to the principal challenges of appearance based models. The main challenges are the recognition of objects in the presence of partial occlusion, the recognition of 3D objects and the classification of objects. The second part of the thesis extends the application of the defined statistical object representation framework to manage these three challenges.

Chapter 1

Introduction and Motivation

During the last few years, there has been a growing interest in object recognition directly based on images, each corresponding to a particular appearance of the object. Representations of objects, which only use 2D–information of images are called *appearance based* models. The interest in such representation schemes is due to their robustness, speed and success in recognizing objects. The benefits of such representation schemes is most obvious in areas like face recognition, human–computer interfaces and content based image retrieval.

This thesis proposes an object representation scheme based on the statistics of robust local neighborhood operators. We want to show that such a representation may provide a robust and highly discriminant means for the recognition of arbitrary objects.

The initial motivation of this work has been the color histogram approach providing fast and robust recognition of colored objects. From an abstract point of view this approach models objects by their color statistics. Our aim has been the generalization of this approach for the modeling of objects by the statistics over their local characteristics. The generalized approach, which we call *multidimensional receptive field histograms*, is able to discriminate a large set of objects in a robust manner. However, the principal defeats are related to what we want to summarize as *challenges for appearance based* models. These challenges are listed below and examined throughout the thesis.

The principal interest of the thesis is therefore the development of a new appearance based representation scheme based on the statistics of vectors of receptive field functions. Different recognition algorithms are proposed. In particular, a probabilistic object recognition algorithm is defined which does not rely on correspondence between testimages and database objects.

Experiments show that high recognition rates can be obtained with different recognition algorithms based on the proposed representation scheme. Extensions of the approach account for the principal challenges which have to be encountered by an appearance based model technique.

In the following we want to motivate the use of a statistical framework (section 1.1) and define object recognition as a case study in computer vision (section 1.2). The initial motivation of the work, the color histogram approach, is briefly discussed in section 1.3. As mentioned above, the principal challenges of appearance based models (section 1.4) will be treated throughout the thesis. Section 1.5 gives an overview of the content of each chapter.

1.1 Motivation for a statistical framework

The thesis proposes the use of a statistical framework for the recognition of objects. Statistics allows to incorporate information into the recognition process such as a priori information, context information and information provided by different sensors. A statistical framework provides a means for the incorporation of such additional information without altering the recognition algorithm itself. In general, adopting a statistical framework may offer the following advantages:

- ability to incorporate uncertainty
- soft decision making
- incorporation of a prior knowledge, which may be independent of the image content

Statistics by its nature makes it possible to incorporate uncertainty. This can be done at different levels as for example at the level of sensor modeling, modeling data incompleteness and decision making. Therefore any doubt or uncertainty can and should be incorporated into such a framework.

Soft decision making refers to the probabilistic character of decisions in the context of statistics. The results of the recognition process may be formulated, for example, as probabilities for each object rather than a hard decision for each object if it is contained in the scene or not. If hard decisions are needed a threshold may be applied ultimately.

We believe that many decisions depend not only on the image content and therefore on the signal itself but also on the context or other a priori knowledge. A statistical framework is particularly suited for the incorporation of such knowledge. Additionally any other source of information may be incorporated as for example the information provided by other sensors.

Whereas statistical modeling and recognition may provide a powerful framework, the main problem of the successful application of statistical algorithms is the fundamental requirement of sufficient data for the estimation of the underlying probability density functions. In this respect the proposed statistical representation bypass the problem by the non–use of positional information. In particular, chapter 4 is devoted to the motivation of the statistical object representation by multidimensional receptive field histograms.

1.2 Object recognition as case study

For our study we have chosen the object recognition problem, since it can be seen as a general case study for computer vision. We can identify different degrees of freedom inherent in the object recognition problem. They are:

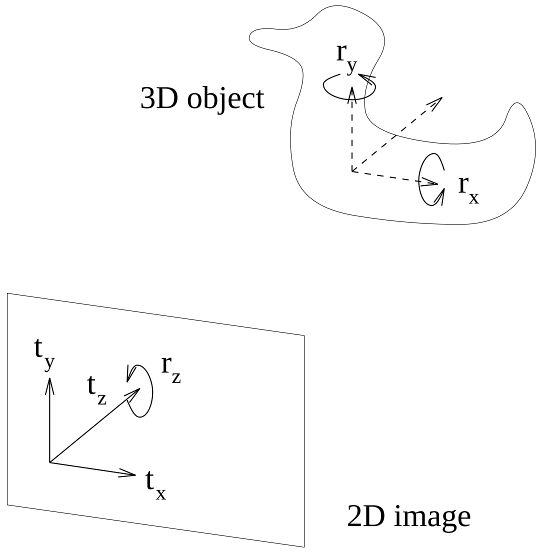


Figure1.1: Different components of rotation and translation of a 3D object

Similarity transformation in the image plane: 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 1.1).

3D transformation of the object: two rotational degrees of freedom (r_x and r_y) exist in addition to the similarity transformation (see figure 1.1).

Scene changes include partial occlusion and background change.

Imaging conditions change accordingly to illumination changes and different types of signal disturbance as signal noise, quantization error and blur. The principal characteristic of these changes is that they cannot be controlled and/or predicted.

Section 4.1 discusses in greater detail the statistical object representation as a function of these degrees of freedom. These degrees of freedom will be examined throughout the thesis. They form, for example, the basis of the structure of chapter 6.

We want to point out the difference between *object identification* and *object classification*, which are both part of the general problem of object recognition. On the one hand side object identification refers to recognition of objects which have been seen by the system before. Object classification on the other hand refers to the capability of the system to generalize outside the database of objects learned and seen before. The second problem therefore is more challenging even though often object identification may be sufficient. In the following we use the term *object recognition* and *object identification* synonymously. The term *object classification* is used for the generalization outside the database.

1.3 Motivation for multidimensional receptive field histograms

Swain and Ballard [Swa 91] have developed a technique which identifies objects in an image by matching a color histogram from a region of the image 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. However, the major drawback of their method is its sensitivity to the color and intensity of the light source and color of the object to be detected. Several authors have improved the performance of the original color histogram matching technique by introducing measures which are less sensitive to illumination changes (see section 2.1).

The color histogram approach is an attractive method for object recognition, because of its simplicity, speed and robustness. Additionally the approach does not rely on the correspondence between the object model and the test image. However, its reliance on object color and light source intensity make it inappropriate for many recognition problems. The initial motivation of our work has been to develop a similar technique using local descriptions of an object's shape provided by a vector of linear receptive fields. For the original color histogram algorithm, it can be seen that robustness to scale and rotation are provided by the use of color. Robustness to changes in viewing angle and to partial occlusion are due to the use of *histogram matching*. Thus it is natural to exploit the power of histogram matching to perform recognition based on histograms of local shape properties. The most general method to measure such properties is the use of a vector of linear local neighborhood operations, or receptive fields.

To this extend, the first part of the thesis (chapters 3 through 6) can be seen as the generalization of color histogram matching to the matching of *multidimensional receptive field histograms*.

1.4 Challenges for appearance based model

Throughout the thesis we will use the term *appearance based model* for techniques, which use 2D–image information for the representation and recognition of objects. The most prominent candidate is based on the principal component analysis of images. But also approaches based on 2D local descriptors and our approach based on the statistics of local neighborhood operators are appearance model based approaches. The main advantages of such techniques is the robustness and speed of the extraction of 2D–image information. By incorporating only 2D–information such approaches may be called *viewer centered* or *image based*. The representation of a 3D–object is therefore based on a collection of 2D–appearance models.

The opposite to appearance based approaches are 3D model based techniques. In such techniques each object has a single 3D model in an object coordinate system. The object models are therefore simpler and less memory intensive as in the case of appearance based models. However, the main disadvantage is the instability of the extraction of 3D–information from 2D–images.

Since the advantages and disadvantages of appearance based models and 3D models are complementary, neither can be stated to be generally superior to the other. Nevertheless, we adopt, as motivated in the previous section, an appearance based approach. This choice is mainly motivated by the robustness and speed which we expect to obtain by the application of such an approach. The *challenges for the application of an appearance based approach* may be summarized and listed as follows:

- recognition in the presence of viewpoint changes
- partial occlusion of objects
- recognition of 3D objects from 2D views
- object classification as the generalization outside the database
- memory requirements of object representation

Throughout the thesis, each point is treated separately. The following sections and chapters can be identified:

- section 6.5 shows the robustness of the object representation by multidimensional receptive field histograms with respect to viewpoints changes
- chapter 7 proposes a probabilistic object recognition algorithm which can account for partial occlusion. Experimental results show that already a small object portion is sufficient for the recognition of 103 objects
- chapter 8 proposes an active object recognition algorithm based on the appearances of objects. The underlying idea is to incorporate several views of an object, resulting in 3D–consistent recognition of objects
- chapter 9 introduces the concept of visual classes as general framework for object classification and proposes a maximum–likelihood recognition of visual classes. Experiments are described for the case of image retrieval
- the memory requirement is analyzed in section 6.7. In the future, memory requirements may be reduced by the application of the concepts for object classification developed in chapter 9 or classical dimensionality reduction techniques

As we can see from the referenced sections and chapters, the principal challenges of appearance based models are the motivation of the second part of the thesis, that is to say for chapters 7 through 9.

1.5 Overview of the thesis

In the following, the content of each chapter is summarized.

Chapter 2 summarizes references which have been source of inspiration for different aspects of the thesis. In particular, we discuss color histogram based approaches. Besides the original approach proposed by Swain and Ballard, several authors have been improving the performance of the approach in the presence of illumination intensity changes. We may point out the popularity of these approaches in the context of image retrieval. Since we generalize color histograms to histograms of vectors of local neighborhood operators, we describe several recognition techniques based on local characteristics. As stated above, we are interested in a general statistical framework so that we describe some statistical object recognition frameworks.

Chapter 3 is devoted to the discussion of local characteristics. Among the most popular characteristics are Gaussian derivatives since they are well understood, can be calculated robustly and they even have a physiological justification. Gabor filter are more time–consuming to calculate in general but widely used in the context of texture analysis. We also describe local descriptor based on color information, which offer the calculation of invariance with respect to changes of the intensity and color of the light source. This chapter also discusses normalization techniques which we apply in order to obtain robustness of local descriptors with respect to noise and illumination intensity changes. The robustness of the proposed normalization techniques with respect to additive Gaussian noise is examined in section 3.2.4. Robustness of the normalization with respect to illumination intensity changes is examined in experiments described in section 5.3.

Chapter 4 develops a general statistical object representation scheme. Each degree of freedom of the object recognition task, as introduced in section 1.2, is discussed and taken into account appropriately. As an approximation of the statistical representation of objects we motivate and propose a set of multidimensional receptive field histograms each corresponding to a particular object appearance. The developed statistical framework may be seen as a general object model for representation schemes based on local descriptor of objects. Based on this general framework we develop an analogy between object recognition and information theory. This allow us to apply information theoretical concepts for object recognition as for example the transinformation for the evaluation of the feature set used.

Chapter 5 introduces different histogram comparison measurements, since the intersection $\cap$ of histograms of the original color histogram approach has limitations in the general context of multidimensional receptive field histograms. The chapter defines and analyzes different histogram comparison measurements as χ^2–statistics, quadratic distances and modified intersection measurements. The computational complexity and characteristics of each measurement are discussed. Additionally, the second part of the chapter analyzes the stability of the measurements with respect to additive Gaussian noise, blurring, image plane rotation and illumination intensity changes.

Chapter 6 applies the introduced comparison measurements for object recognition by histogram matching. After a recognition example for 261 objects, different degrees of freedom of the object recognition task are taken into account. In particular, the chapter describes how to account for image plane rotation and scale changes and provides experimental results in the presence of such changes. Section 6.5 is devoted to the important aspect of object recognition in the presence of viewpoint changes and analyzes the robustness of multidimensional receptive field histograms with respect to such changes. At the end of the chapter the memory requirement for multidimensional histograms is discussed.

Chapter 7 extends the application of multidimensional receptive field histograms to probabilistic recognition of objects. The probabilistic algorithm is capable of recognizing objects from a small portion of the image providing robustness to partial occlusion. Recognition results are given for a database of 103 objects in the presence of image plane rotation, scale changes and viewpoint changes. Based on these results we propose a "dynamic hashtable" approach using image patches as index of the hashtable. This latter algorithm is particularly suited for the recognition of objects in cluttered scenes.

The following two chapters propose two further extensions and applications of multidimensional receptive field histograms for active recognition of objects and for object classification. The experimental results provided in these chapter are not complete so that these chapters may be seen as perspective of the thesis.

Chapter 8 adopts a hypothesis–testing strategy for the active recognition of objects in a single 2D–image as well as in 3D. For the case of 2D–images we develop a general interest point detector which may be applied also in the context of other object recognition algorithms. The second active recognition algorithm uses the information theoretic concept of transinformation for the evaluation of the most salient viewpoints of an object. Moving the camera to these salient viewpoints make it possible to verify an object hypothesis derived from another viewpoint. Since at each viewpoint only 2D–information is used for recognition, this algorithm enables an appearance based model approach to recognize 3D–objects from several 2D–views in a consistent way. Experimental results underline this property of the algorithm.

Chapter 9 proposes the concept of visual classes as a general framework for object classification based on results of previous chapters. Visual classes are defined on 2D– and/or 3D–similarities of objects which may be derived from the general statistical representation of objects. The chapter proposes a maximum–likelihood approach for the recognition of visual classes and applies the technique for the retrieval of visually similar images.

Chapter 10 concludes the principal results and lists perspectives of the thesis.

Chapter 2

State of the art

This chapter summarizes briefly some references which have been sources of inspiration for this thesis. Here, we do not aim to give a comprehensive overview of the numerous object recognition algorithms (see for example [Obj 96, Pop 95, Gri 92, Gri 91]). Throughout the thesis we are particularly interested in object models which can be obtained or estimated automatically from a set of images of an object.

As stated in the introduction, we have been inspired initially by the simplicity, speed and robustness of the color histogram technique which is described in the first section of the chapter. In order to generalize the color histogram approach, color will be replaced by multidimensional receptive field vectors or vectors of local neighborhood descriptors. Section 2.2 therefore summarizes four object recognition algorithms based on vectors of local descriptors. Due to the importance of the choice of local characteristics we devote chapter 3 to this topic. Seen from a more abstract point of view, a histogram of vectors of local neighborhood descriptors is an approximation of the probability density function over the vectors. The last section 2.3 therefore summarizes three statistical object recognition algorithms.

2.1 Color histogram based approaches

The next section 2.1.1 introduces the color histogram approach [Swa 91] and some extensions which aim for illumination independence. Section 2.1.2 describes an approach where geometric and color information is combined. The last section 2.1.3 points out importance of color histograms in the context of content based image retrieval.

2.1.1 Color indexing

Swain and Ballard [Swa 90, Swa 91, Swa 93a] have proposed a technique where each object is represented by its color histogram (an approximation of its color distribution). Objects are identified in an image by matching a color histogram from a region of the image with a color histogram from a sample of the object (see chapter 5 for histogram comparison measurements). 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. However, the major drawback of their method is its sensitivity to the color and intensity of the light source. Furthermore, not all object classes can be described by color alone. Several authors have improved the performance of the color histogram approach by introducing measures which are less sensitive to illumination changes (see below and [Fun 95, Hea 94, Enn 95]).

The color histogram approach is an attractive method for object recognition, because of its simplicity, speed and robustness. Additionally, the use of color histograms for recognition does not assume segmentation of an object nor an explicit geometric model. An object is described only by its color histogram. However, its reliance on object color, light source intensity and light source color make it inappropriate for many recognition problems. For the Swain and Ballard algorithm, it can be seen that robustness to scale and rotation are provided by the use of color. Robustness to changes in viewing angle and to partial occlusion are due to the use of *histogram matching*. Thus it is natural to exploit the power of histogram matching to perform recognition based on histograms of local shape properties. The most general method to measure such properties is the use of a vector of linear local neighborhood operations, or receptive fields.

The sensitivity of the color histogram approach to light intensity and also to light color makes the choice of the histogram axis crucial. Swain and Ballard propose to use a variant of the "opponent" color axis which is given by the following equation [Bal 82, Swa 90]:

$$\begin{pmatrix} rg \\ by \\ wb \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & -1 & 2 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.1}$$

This linear transformation of the RGB-space is typically advocated as the way humans perceive color. In the context of color histograms the linear transformation allows discretization of the "intensity" wb more coarsely than the other two axis. The proposed resolution of the axis is 8 bins for the wb-axis and 16 for the remaining two, which makes a total of 2048 cells per histogram. Nevertheless the linear transformation does not reduce the sensitivity to color intensity significantly. Swain and Ballard therefore propose to use a color constancy algorithm prior to histogramming in order to reduce the sensitivity to illumination changes.

As mentioned above, several authors have improved the original approach in order to reduce the sensitivity to illumination intensity changes. The two most successful approaches have been proposed by Healey/Slater [Hea 94] and Funt/Finlayson [Fun 95]. Healey and Slater calculate moment invariants of the entire color histogram. The invariants are based on the finite dimensional linear color model, which allows the modeling of illumination intensity changes by a linear transformation between color histograms. Even though the experimental results [Hea 94] are convincing the method is global in the sense that the intensity change is assumed to be constant over the entire image. This makes the approach inappropriate in many situations [Col 96].

Funt and Finlayson [Fun 95] use derivatives of the logarithms of the color channels in order to provide illumination invariant features. The underlying assumption is a local constant illumination. The color invariants are based on an approximative color model (coefficient model) which allows us to calculate color invariants locally. Even though the underlying model is a simplification these invariants perform best when compared with other color invariants [Col 96] as for example proposed by Nagao [Nag 95]. The main advantage of the proposed invariants is that they can be calculated locally and do not assume a uniform illumination change in the entire image. It is worth mentioning that histograms of the proposed invariants do not code color information itself but rather relations between adjacent color regions.

An interesting extension of the original color indexing method has been proposed by Ennesser and Medioni [Enn 93, Enn 95]. By using local color histograms of a test image their algorithm is able to extract regions of the image which are most likely to contain a particular object. Experiments show that the algorithm is able to detect objects in relatively complex environments.

Hunke, Schiele and Waibel [Hun 94, Sch 95c, Sch 95b] have shown that normalizing the color vector by luminance provides a reliable means to detect skin color for tracking human faces. They exploit the simplicity of the approach to design a system for real–time tracking of human faces under varying conditions.

2.1.2 Color and geometric information for indexing

Several authors have combined the color histogram approach with some type of geometric information in the context of recognition. Examples include [Sla 95, Mat 95]. This section describes one such system, neurally inspired, called SEEMORE [Mel 96], which is of particular interest.

SEEMORE uses 102 different feature channels. The response of each individual feature channel is calculated by sub-sampling and summing the output of the element channel feature over the entire visual field (or more precisely over the pre-segmented region of the image). The 102 sums over the feature channels are the principal input for a nearest–neighbor classifier, which is trained on sample images. The feature channels fall into five categories: 23 circular color channels (described by hue and saturation), 11 coarse–scale intensity corner channels, 12 circular and oriented "intensity blob channels", 40 generalized contour shape channels (as junctions, curves, parallel and oblique contour pairs) and 16 oriented Gabor–derived texture channels (using Energy and relative orientation at different scales and orientations).

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. Unfortunately the color representation by hue and saturation generally needs constant lighting conditions so that the algorithm does not perform well in the presence of illumination changes. Additionally the 102 feature channels were chosen by hand.

2.1.3 Image retrieval

Recently one of the most active field of research has been content based image retrieval of large video and/or image databases. Research groups at MIT [Pen 96, Pic 96], Stanford [Tom 94, Gui 96], University of California Berkeley [Bel 97] and IBM [Fli 95] are among the most active groups. Typically, color is used among other clues for indexing into large databases. The system of Berkeley and the system of IBM directly use color histograms for indexing. Picard at MIT also proposes a histogram technique for the sorting of images at "a glance" [Gor 94, Pic 96].

The interest in color is derived from the fact that many images contain characteristic colors. The interest in color histogram techniques is also motivated by the low computational cost and the relatively low storage requirement. Thus it will be interesting to analyze multidimensional receptive field histograms as a generalization of color histograms in the context of image retrieval.

2.2 Object recognition based on local characteristics

The section briefly describes four approaches which use local descriptors for recognition. The next section 2.2.1 describes the geometric hashing approach based on point features but which can be extended to more general features. The next section 2.2.2 describes a system developed at the University of Rochester for real–time recognition by an iconic sparse distributed memory. Section 2.2.3 describes an algorithm which profits from the strengths of the previous two. This system is certainly one of the most successful object recognition systems to date. The last section 2.2.4 shortly introduces a two stage memory based recognition system which can be seen as an efficient implementation of the general Hough transform.

2.2.1 Geometric hashing

Lamdan et al. [Lam 88c, Lam 88b] describe a general framework for recognition which is particularly suitable in scenes with overlapping and partially occluded objects. Models are described by sets of interest points which are invariant under an affine transformation. In order to reduce the time and the complexity of recognition all possible combinations of interest point sets are stored in a hash–table. During recognition sets of interest points are extracted from the scene and used for indexing into the hash–table and voting for object models. Recognition therefore becomes a point (or in more general a feature) matching task.

The first step of the algorithm is an interest points operator, however its choice is not essential for the rest of the algorithm, since only the coordinates of the interest points are used in what follows. Assuming the extraction of m interest points a non–collinear triplet of interest points is chosen as an affine basis. The coordinates of the remaining $m-3$ interest points are calculated and used as entry to the hash–table (together with the associated object label). In order to reduce the complexity of recognition all possible non–collinear triplets are used as affine basis. The overall complexity of this first (off-line) step is in the order of $O(m^4)$.

During recognition the same interest point operator is applied for the extraction of n interest points. An arbitrary non–collinear triplet is used as affine basis and the affine coordinates of the remaining $n-3$ points are used for voting for object models (or more precisely for a particular triplet basis and its object model). The complexity of a single voting step is $O(n)$. If none of the object models obtains a sufficient number of votes another non–collinear triplet can be used for a second voting step. The worst case complexity (for example rejecting a test image) is $O(n^4)$ using all possible triplets. When sufficient votes are accumulated for a particular object model triplet it can be verified by a least square match between the test image and the transformed object model [Lam 88c]. ([Lam 88b] generalizes the approach to similarity and projective transformation: in the first case only two points are needed to define a basis and for the projective case 4 points are necessary for the definition of a projective basis. The resulting complexity is then $O(m^{k+1})$, with $k=2$ and $k=4$ respectively.)

The most interesting property of the algorithm is that the underlying model representation is entirely *local*: a set of interest points is by definition local (only a small portion of the image is used). Since an object model consists of a collection of interest point sets, the model itself is local. An interest point set can be seen as a high–level feature which is used for indexing. The algorithm can be seen therefore as a feature matching algorithm.

Another advantage of the approach worth mentioning is that the voting not only includes the object identity but also the location of the object in the scene (in means of the basis triplet).

Several attempts have been made in order to reduce the complexity (or at least the expected recognition time) of the algorithm. The obvious dilemma is that the more points are used for the construction of a set the more distinctive the sets become, but the more the complexity increases. Using k-point sets the complexity becomes $O(m^{k+1})$ for the learning respectively preprocessing step and $O(n^{k+1})$ for the worst case of recognition. This dilemma is due to the fact that only (affine) point coordinates are used without the use of the local structure around the point (or the classification of the point).

In [Lam 88a, Wol 90] the original interest points are replaced by interest features which provide more information. Nevertheless such features have to be invariant under the considered transformation (for example affine transformation) which limits the feature choice.

[Gri 90] provides a theoretical analysis of the sensitivity of the geometric hashing technique. The main result is that the probability of false positives (during voting) increase considerably already in the presence of moderate noise in the data points. An improved probabilistic voting scheme is provided by [Rig 93]. More recent work includes [Beb 95] and [Lam 96].

2.2.2 Iconic sparse distributed memory

The basic principle of an approach proposed by Wixson, Rao and Ballard [Bal 93, Rao 95b, Rao 95a] is the representation of objects (or object patches) by a high–dimensional "iconic" feature vector. Rao and Ballard [Rao 95b] stress in particular the favorable matching properties of high dimensional spaces in high precision recognition, since they provide most of the time a unique response for a particular object. Following an argument of Karneva, [Rao 95a] claims that a high–dimensional iconic feature vector of a particular object can be subjected to considerable noise before it is confused with vectorial representations of other objects.

The employed high dimensional feature vector $r(x, y)$ (centered at the image location (x, y)) includes 45 responses of nine oriented Gaussian filters at five different scales $(9 \times 5 = 45)$:

$$r(x,y) = \begin{pmatrix} r_{\sigma_1}(x,y) \\ r_{\sigma_2}(x,y) \\ r_{\sigma_3}(x,y) \\ r_{\sigma_4}(x,y) \\ r_{\sigma_5}(x,y) \end{pmatrix} \text{ with } r_{\sigma_i}(x,y) = \begin{pmatrix} G^{\sigma_i}_{1,0} \\ G^{\sigma_i}_{1,\frac{\pi}{2}} \\ G^{\sigma_i}_{2,0} \\ G^{\sigma_i}_{2,\frac{\pi}{3}} \\ G^{\sigma_i}_{2,\frac{2\pi}{3}} \\ G^{\sigma_i}_{3,0} \\ G^{\sigma_i}_{3,\frac{\pi}{4}} \\ G^{\sigma_i}_{3,\frac{\pi}{2}} \\ G^{\sigma_i}_{3,\frac{3\pi}{4}} \end{pmatrix} \star I(x,y) \quad i = 1, 2, \ldots, 5 \tag{2.2}$$

In order to make the vector $r(x, y)$ invariant to image plane rotation the vector response is normalized: selecting the maximum orientation of the first Gaussian derivative as reference orientation, all components of the vector are steered into this direction (by employing the steerability property of Gaussian derivatives, see section 3.1.1 for details).

For object indexing the feature vectors $r(x, y)$ from different objects are stored in a generalized version of Karneva's sparse distributed memory, which is suitable for the large address space spanned by the 45 dimensional feature vectors. During learning and object recognition a figure ground segmentation is performed (by using zero cross disparity filtering of stereo image pairs). The vector responses are calculated on a MV200 pipeline image–processing system which allows the calculation of convolutions at frame rate. The memory access is also implemented on the MV200 system which speeds up recognition.

One drawback of the proposed feature vector $r(x, y)$ is its relatively large support: the responses of the vector $r(x, y)$ 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 $r(x, y)$ is in the order of 128×128 pixels. The vector $r(x, y)$ therefore cannot be called local, since it already covers a $\frac{1}{16}$ of a typical 512×512 image. In order to calculate $r(x, y)$ more locally the support of the vector has to be reduced, which would compromise on the uniqueness of the filter response.

Due to the global character of the feature vector, the approach is sensitive to occlusion. [Bal 94] introduces a separate algorithm which can account for partial occlusions. The algorithm relies on the fact that a figure ground separation can be achieved (for example by zero-disparity between an stereo image pair). The basic idea is to reconstruct an image patch approximately by a pseudo inverse transformation from a single feature vector $r(x, y)$. By masking the occluded parts of the reconstructed image patch one can calculate the expectation vector $\hat{r}(x, y)$ relative to the estimated occlusion in the image. This estimated vector $\hat{r}(x, y)$ can be compared to the observation in the image.

2.2.3 Recognition as correspondence problem

One of the most successful object recognition algorithms has been proposed by Schmid and Mohr [Sch 96h, Sch 96i]. The algorithm consists of three major steps: interest point detection, characterization of interest points by vectors of local descriptor and storage of each vector in a hash table. 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 interest point detector is a modified version of the "Harris-detector" and reduces the data from the entire image to a number of interest points. Each interest point in an image is described by a nine–dimensional rotational invariant vector of local characteristics based on Gaussian derivatives. These derivatives were proposed by Koenderink [Koe 87] and use derivatives up to the third order. Finally the vector responses of all interest points of an image are stored in a hash table indexed by the nine–dimensional vector.

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 considered 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). By using not only the vector responses but also some geometrical invariants between different interest points the voting scheme becomes more selective. Another possibility for improvement is the use of a probabilistic voting scheme, recently proposed by Mohr et al. [Moh 97].

Impressive experimental results have been presented on a database of several hundred objects. Nevertheless, probably the weakest point of the approach is the application of an interest point detector. The success of the approach relies on the the repeatability of the interest point detector over different image and viewing conditions (as for example illumination changes and scale changes), which is difficult to achieve. Furthermore the approach, as presented in [Sch 96h], is optimized for a fast access of a single interest point (hash table). That implies that the representation of an image (object) is not directly accessible (without scanning the entire hash table). This makes it difficult to generalize from the stored images and to find global similarities between different objects and/or images.

2.2.4 Two stage memory based recognition

Nelson [Nel 95, Nel 96] proposes a method, that combines an associative memory with a Hough–like evidence combination technique. The method is based on robust (Nelson calls them semi–invariant) local *keys* associated with object hypothesis, evidence and object configuration. The main motivation is the generality and the relative simple learnability of memory based representations from visual data.

The method utilizes an associative memory organized so that access via a key feature evokes associated hypotheses for the identity and configuration of all objects that could have produced it. The key features therefore have to be robustly extract-able and should contain sufficient information to specify a configuration of an object. The second stage of the method uses a second associative memory which is indexed by the configuration (and identity) of objects. This second associative memory maintains a probabilistic estimate of the likelihood of each hypothesis and the statistics about the occurrence of the keys in the primary memory (the statistics are mainly used to improve the voting algorithm). The idea is similar to a multidimensional Hough transform but reduces considerably the degrees of freedom of the considered transformation space.

The choice of key features for this approach is particularly important: key features should be local in order to be robust to occlusion and clutter. Robustness of the key features helps to reduce the storage requirement. Additionally, keys should be discriminant but at the same time robustly extract-able. In particular the last two properties (robust detect-ability and ability to discriminate) complicate the choice of key features. Nelson proposes the construction of higher level (discriminant) features based on structural grouping of simple features (equivalent to perceptual grouping). Segment chains of three connected segments are used [Nel 95] for the representation of polyhedral objects. [Nel 96] additionally uses curve patches for indexing, which by definition contains segment chains as a special case.

Experiments reported in [Nel 96] show a performance of the algorithm between 70% and 95% for a database containing 7 objects, depending of the degree of occlusion and clutter. The reason for failure is typically the poor performance of the low–level feature identification. Even with robust features, the number of viewpoints per object (for the whole viewing sphere) was in the order of 100 (which corresponds to the whole viewing sphere sampled every 20 degrees and without any symmetry of the object).

2.3 Statistical object recognition

Statistical methods are frequently used in computer vision. Typical application areas are the segmentation of images [Can 86, Bel 89] and the probabilistic modeling of the correspondence and matching problem [Bre 93, Hut 95]. Even though statistical models are also applied in the context of object localization and recognition, objects themselves are rarely modeled by statistical density functions. In the context of this thesis we are mainly interested in the automatic acquisition of a statistical object model.

Three approaches are briefly described in the following. They have in common that a statistical object model is learned from image data. The eigenpicture approach (section 2.3.1) directly uses the distribution of image pixel values. Section 2.3.2 describes the application of the Expectation–Maximization algorithm for the probability density estimation of point features. Section 2.3.3 finally describes the unsupervised learning of a probabilistic graph of objects.

One major difficulty with the statistical learning of object density functions is the lack of sufficient image data. Since Artificial neural networks are in principal able to generalize outside the training set, some of the most successful recognition systems are based on such techniques [Pog 90]. [Shv 90, Ede 93] also provide theoretical analysis of the learnability of object models from images. Artificial neural networks are not further considered throughout the thesis.

2.3.1 Eigenpicture approaches

Recently many researchers [Sir 87, Kir 90, Tur 91a, Tur 91b, Mur 93, Mur 95, Mog 95, Ohb 96] use the Karhunen–Loeve transformation [Fuk 90] for the calculation of *eigenpictures* in the context of object recognition. The general idea is to interpret an image with $n = n_1 \times n_2$ pixels as one single n–dimensional vector ϕ. Having L images interpreted as n–dimensional vectors $\phi_i (i = 1, 2, \ldots L)$, the covariance matrix C is given by (with $\overline{\phi} = \frac{1}{L}\sum_{i=1}^{L}\phi_i$):

$$C = \frac{1}{L}\sum_{i=1}^{L}(\phi_i - \overline{\phi})(\phi_i - \overline{\phi})^T \tag{2.3}$$

Let's assume the calculation of the eigenvectors u_i with their corresponding eigenvalues λ_i (see [Fuk 90, Pre 92] and [Tur 91a] for an algebraic reformulation which reduces the complexity of the eigenvector calculation considerably) and that for all $i = 1, 2, \ldots L-1 : \lambda_i \geq \lambda_i + 1$. An optimal (in the least square error sense) K–dimensional basis for the representation of the L images is given by the first K eigenvectors $u_1, u_2, \ldots, u_K$. Even though the Karhunen–Loeve transformation (or principal component analysis) is not new in the pattern recognition community it has been shown to be one of the most successful recent approaches for object recognition.

In the context of face recognition Turk and Pentland [Tur 91a] and more recently Pentland et al. [Pen 94, Mog 95] have applied the eigenpicture technique successfully to large face databases. A real–time 3D–object recognition system has been implemented by Murase and Nayar [Mur 95] based on a "modular eigenspace".

The main advantage of the 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 example 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. The second major drawback of the approach is that the modeling of each image is global, which makes the approach sensitive to partial occlusion.

The calculation of fisher-faces (instead of eigenfaces) has been proposed by Belhumeur et al. [Bel 96]. Their approach performs better in the presence of illumination intensity changes than the original approach. [Ohb 96] calculate eigenpictures of small image patches which makes the approach more suitable in the presence of clutter and partial occlusion. Eigenpictures have been also applied for tracking of 3D–objects [Bla 96], gesture recognition [Mar 97] and transmission of images for video conferencing [Cro 96]. Overall the eigenpicture approach has been shown to be successful in several areas of computer vision.

2.3.2 Statistical learning, localization and identification of objects

Hornegger and Niemann [Hor 95, Hor 96] propose an object representation by parameterized density functions of object features. Both the learning as well as the localization of objects are formalized as a parameter estimation task. In the proposed framework, features are interpreted as random variables. An object is therefore represented by the probability density function of its set of features. In [Hor 95] point coordinates are used as features. Any transformation of the object (as for example rotation, translation, scale change, viewpoint change) are represented by a parameter of the probability density function. The authors consider, in particular, the problem of 3D object recognition from 2D views which adds the projection from model space into the image plane. Or *vise versa* the authors encounter the problem of estimation the 3D model (estimation of the probability density function of a 3D object) from 2D images.

During an unsupervised training stage (no correspondence is assumed) the probability density function of a 3D object is estimated from 2D images by using the Expectation–Maximization algorithm. This algorithm is particularly useful for this type of incomplete estimation problem. The probability density function is represented by a mixture of multivariate Gaussian density functions. This modeling is reasonable in the proposed context of point features. Unfortunately the generalization of the method to more general feature sets is not reported.

Even though the proposed approach is sound, only some experimental results are provided. In [Hor 95] only four 2D objects and two 3D objects are used. Additionally, the recognition time is relatively slow, due to the fact that the identification procedure is formulated as a parameter estimation problem (optimization problem).

2.3.3 Automatic model acquisition

Several authors [Pop 93, Bei 94, Beb 95, Pop 95, Pop 96] are particularly interested in the automatic acquisition of object models for recognition. The two approaches reported here follow the same philosophy: based on simple features (as for examples lines, edges) higher level features are constructed by perceptual grouping. Based on such features and feature sets, [Bei 94] proposes to learn a probabilistic hash table for indexing. [Pop 93] proposes to learn the relations between the different features. An object model (as well as a test image) is represented by a graph, where the nodes represents features and arcs represent the spatial relationship between the features. In particular the probability of observation of a feature is learned, which is used during recognition in a probabilistic alignment scheme. It is interesting to note that the probability density function is obtained by a non–parametric estimation approach (Parzen–estimate [Fuk 90]). For each 2D–aspect of an object such a probabilistic graph is learned automatically. A 3D object is therefore represented by a collection of probabilistic graphs.

As in the previous case of section 2.3.2 the approach has been applied only on a relatively small set of objects.

2.4 Conclusion

Without giving a comprehensive overview, the chapter summarized object recognition algorithms which are of special interest for the thesis. Three types of approaches have been emphasized: color histogram based approaches, since their speed and robustness were the initial motivation for the thesis. Approaches based on local characteristics are of interest, since we want to use statistics of local characteristics for the representation of objects. Statistical recognition approaches have been described, showing benefits by adopting a statistical framework for object recognition.

The next chapter is devoted to the discussion of local characteristics as Gaussian derivatives, Gabor filters and color information. The reminder of the thesis uses statistics of vectors of such local descriptors for the representation of objects.

Chapter 3

Local descriptor

All object recognition systems rely – explicitly or implicitly – on some type of feature. The choice of the used features is critical and depends on such factors as the considered object classes, the sensor characteristics, the context and the task. Generally, there exists a tradeoff between precision and generality of the features. Even though researchers are aware of the difficulties, the choice is most often made by hand and arbitrarily. One contribution of the thesis is a qualitative measurement that enables the comparison of different feature sets relative to a given object set (see the application of information theory to object recognition in section 4.2.3).

The thesis proposes a general framework of object representation as a probability density function over a set of local descriptors of the object appearance (see section 4.1). In order to be as general as possible, the used local descriptors are not restricted to a particular family of objects or to a particular type of descriptor. Nevertheless, it is necessary to formulate some minimal requirements for local descriptors. One basic requirement is the *locality* of the features. In contrast to local features, global features are sensitive to partial occlusion and to any local disturbance of the image (for example specular reflections). Another basic requirement is the *robustness* of features. Robustness is a weaker assumption than invariance which is difficult to achieve for the general case. We can list three categories of features:

invariant features: such features are considered constant with respect to a certain transformation (examples include affine and orthographic transformations),

equivariant features: the feature value is a function of a certain transformation,

robust features: the feature value changes slowly in the presence of some transformations.

The invariance of features is very attractive property, since it reduces the set of possible feature values. In general we should use invariant features whenever practicable (in the sense of stability and ability to discriminate). Unfortunately invariant features typically restrict the object classes. Often they are based on the calculation of higher order derivatives (instability problems) and/or they are global (problems with partial occlusion). Either of these constraints would limit the generality of our approach. Consequently we would like to "weaken" the requirement of invariance.

Equivariant features are given for example by Gaussian derivatives with respect to image plane rotation and scale changes. Nevertheless this property cannot generally be achieved. More general is the notation of robust features. Our claim is that many local descriptors, which are suitable for object recognition, can be calculated in a robust manner so that they contain discriminant information. In the following we are mainly interested in robust features (and if possible equivariant features) which can be calculated locally and robustly with respect to image noise, image blur, image plane rotation and scale changes.

Section 3.1.1 introduces Gaussian derivatives, their steerability with respect to image plane rotation and the equivariance property to scale changes. 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 95b]), their capacity to model the response of neural cells [You 86] and the existence of a recursive implementation [Der 93]. Furthermore Gaussian derivatives (as well as Gabor filters) are robust to scale changes of approximately $\pm 20\%$ [Sch 96h]. The experiments described throughout the thesis mainly use Gaussian derivatives because of their robustness as well as because of the steerability to image plane rotation and the equivariance property to scale.

Gabor filters (section 3.1.2) satisfy nearly the same constraints as Gaussian filters (robustness, steerability to image plane rotation, equivariance to scale changes). We used Gabor filters in early experiments. Since the recognition results with Gabor filters and with Gaussian filters are almost identical we do not describe these experiments in this report.

Color has been examined in a study related to this thesis [Col 96]. The study showed in particular that the histograms of the derivatives of the logarithms of color channels provide a means of recognition in the presence of changes of illumination intensity and even in the presence of light source color shifts. A short resume of the theoretical basis of these invariants is provide in section 3.1.3.

Section 3.2 describes normalization procedures which can be used in order to make the output of particular filters more robust to Gaussian noise and illumination intensity changes. The robustness of the normalization techniques to additive Gaussian noise is reported in section 3.2.4. The robustness to illumination intensity changes is examined later in section 5.3.

3.1 Local characteristics

Section 3.1.1 describes Gaussian derivatives their steerability to image plane rotation and their equivariance to scale changes. Gabor filters and a particular choice of the filter parameter are briefly introduced in section 3.1.2. Section 3.1.3 discusses some invariants based on color information.

3.1.1 Gaussian derivatives

This section introduces local characteristics based on Gaussian derivatives. Gaussian derivatives are widely used in the literature and well understood [Fre 91, Rao 95b]. 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 $G^\sigma(x,y)$:

$$G^\sigma(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3.4}$$

the first order derivative in the x–direction is given by:

$$G_x^\sigma(x,y) \quad = \quad -\frac{x}{\sigma^2}G^\sigma(x,y) \tag{3.5}$$

More generally the first order derivative in direction $\vec{v} = (\cos\phi \ \sin\phi)^T$ is given by:

$$G_{1,\phi}^\sigma(x,y) = \frac{\partial}{\partial\vec{v}}G^\sigma(x,y) \tag{3.6}$$

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

$$G_{n,\phi}^\sigma(x,y) = \frac{\partial^n}{\partial\vec{v}^n}G^\sigma(x,y) \tag{3.7}$$

In this thesis we use only 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:

$$G_x^\sigma(x,y) \quad = \quad G_{1,0°}^\sigma(x,y) = -\frac{x}{\sigma^2}G^\sigma(x,y) \tag{3.8}$$

$$G_y^\sigma(x,y) \quad = \quad G_{1,90°}^\sigma(x,y) = -\frac{y}{\sigma^2}G^\sigma(x,y) \tag{3.9}$$

Based on these first order derivatives one can define the magnitude $Mag(x,y)$ and the direction $Dir(x,y)$ of the first derivative:

$$Mag(x,y) \quad = \quad \sqrt{(G_x^\sigma(x,y))^2 + (G_y^\sigma(x,y))^2} \tag{3.10}$$

$$Dir(x,y) \quad = \quad \arctan\frac{G_y^\sigma(x,y)}{G_x^\sigma(x,y)} \tag{3.11}$$

Three second order derivatives are given by:

$$G^{\sigma}_{xx}(x,y) = (\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2})G^{\sigma}(x,y) \tag{3.12}$$

$$G^{\sigma}_{xy}(x,y) = (\frac{xy}{\sigma^4})G^{\sigma}(x,y) \tag{3.13}$$

$$G^{\sigma}_{yy}(x,y) = (\frac{y^2}{\sigma^4} - \frac{1}{\sigma^2})G^{\sigma}(x,y) \tag{3.14}$$

Based on the first and second order derivatives one can define rotation invariant characteristics:

$$Lap(x,y) = G^{\sigma}_{xx}(x,y) + G^{\sigma}_{yy}(x,y) \tag{3.15}$$

$$G12(x,y) = G^{\sigma}_{xx}(G^{\sigma}_{x})^2 + 2G^{\sigma}_{xy}G^{\sigma}_{x}G^{\sigma}_{y} + G^{\sigma}_{yy}(G^{\sigma}_{y})^2 \tag{3.16}$$

$Lap(x,y)$ is the well known Laplacian operator. The second local characteristic will be called $G12(x,y)$ throughout the thesis, since it is based on first and second order derivatives. $G12(x,y)$ was introduced by Koenderink [Koe 84] as a rotation (image-plane rotation) invariant filter and has been used by Schmid and Mohr [Sch 96i, Sch 96h] for object recognition. Three rotation invariant filters, which are based on first and second order Gaussian, are used: $Mag(x,y)$, $G12(x,y)$ and $Lap(x,y)$ (equations 3.10, 3.16 and 3.15 respectively).

In the following we introduce the equivariance of Gaussian derivatives to scale and the steerability to image plane rotation.

Equivariance of Gaussian derivatives to scale

As mentioned above local characteristics should be calculated at an arbitrary scale. This is possible not only for Gaussian derivatives but also for other families of filter as for example Gabor filters (see section 3.1.2). In the following we introduce the equivariance property of Gaussian derivatives to scale.

Given a two–dimensional function $p(x,y)$ and its scaled version $f(x,y) = p(sx,sy)$ we know from analysis:

$$f(x,y) = p(sx,sy) \tag{3.17}$$

$$\frac{\partial}{\partial x}f(x,y) = s\frac{\partial}{\partial x}p(sx,sy) \tag{3.18}$$

$$\vdots$$

$$\frac{\partial^n}{\partial x^n}f(x,y) = s^n\frac{\partial^n}{\partial x^n}p(sx,sy) \tag{3.19}$$

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 on the basis of the $p(x,y)$. In the following we show this property for the first order derivative. The first order derivative of f is defined as:

$$\frac{\partial}{\partial x} f(x,y) = G_x^\sigma(x,y) \star f(x,y) \tag{3.20}$$

where $G_x^\sigma(x,y)$ is the Gaussian derivative (see equation 3.8) and $\star$ is the convolution operator. Therefore we obtain (together with equation 3.18):

$$\begin{aligned} \frac{\partial}{\partial x} f(x,y) &= s\frac{\partial}{\partial x} p(sx,sy) & (3.21)\\ &= sG_x^\sigma(x,y) \star p(sx,sy) & (3.22)\\ &= sG_x^{\sigma s}(x,y) \star p(x,y) & (3.23) \end{aligned}$$

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 *adaption of the Gaussian derivative to scale.* In a similar way one obtains an equation for the adaptation of the nth order derivative to scale:

$$\frac{\partial^n}{\partial x^n} f(x,y) = s^n G_{x^n}^{\sigma s}(x,y) \star p(x,y) \tag{3.24}$$

Following this equation, one can calculate the nth order derivative of a function $f(x,y)$ directly from the basis of 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, one has to adapt the support for the calculation of the nth order derivative of p. This is expressed by the adaptation of the standard deviation σs of the Gaussian filter.

We call the adaption of the Gaussian derivatives to scale changes by the factor s the *equivariance* property of the Gaussian derivatives to scale. More generally we call any local characteristic equivariant to a particular change, whenever there exist a certain parameter which is directly connected with this change.

As expected, the *equivariance to scale* is not only true for characteristics based on Gaussian derivatives. The same property holds, for example, for Gabor filters due to their Gaussian envelope. Gabor filter are introduced in section 3.1.2.

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 one of the filter responses of many different orientations have to be calculated, it is very time consuming to calculate every time the corresponding version of the filter. Therefore it is desirable to define a finite set of basis filters and an interpolation rule, which allows us to calculate the filter response based only on the response of the basis set. For the first order Gaussian derivative there exists a well known linear interpolation rule [Fre 91]:

$$G_{1,\phi}^\sigma = \cos\phi G_x^\sigma + \sin\phi G_y^\sigma \tag{3.25}$$

In order to formalize this property, Freeman and Adelson [Fre 91] start with a two–dimensional function $F(x,y)$ and define $F^{\theta}(x,y)$ as the version of $F(x,y)$ rotated by the angle θ. A function $F(x,y)$ *steers* when it can be written as a sum of rotated versions of itself:

$$F^{\phi}(x,y) = \sum_{j=1}^{J} k_j(\phi) F^{\theta_j}(x,y) \tag{3.26}$$

This equation is known as the steering constraint. J corresponds to the number of interpolation functions $k_j(\theta)$, the F^{θ_j} form the finite basis set of oriented functions. Two important questions arise: how many interpolation functions are needed and how to obtain the interpolation functions. In order to formulate two theorems, the function $F(x,y)$ is rewritten in polar coordinates $r = \sqrt{(x^2+y^2)}$ and $\rho = \arg(x,y)$:

$$F(r,\rho) = \sum_{n=-N}^{N} a_n(r) e^{in\rho} \tag{3.27}$$

A first theorem [Fre 91] states that the steering constraint 3.26 holds for functions expandable in the form of 3.27 if and only if the interpolation functions $k_j(\phi)$ are solutions of:

$$\begin{pmatrix} 1 \\ e^{i\phi} \\ \vdots \\ e^{iN\phi} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ e^{i\theta_1} & e^{i\theta_2} & \cdots & e^{i\theta_J} \\ \vdots & \vdots & \ddots & \vdots \\ e^{iN\theta_1} & e^{iN\theta_2} & \cdots & e^{iN\theta_J} \end{pmatrix} \begin{pmatrix} k_1(\phi) \\ k_2(\phi) \\ \vdots \\ k_J(\phi) \end{pmatrix} \tag{3.28}$$

By fixing θ_j the interpolation functions are given as solutions of equation 3.28. A second theorem [Fre 91] states that the minimal number of interpolation functions is T, whereas T is the number of nonzero coefficients $a_n(r)$ in equation 3.27. Therefore the theoretical answers to the two questions are known: how many interpolation functions are needed and how are the interpolation functions are defined.

The two theorems can be applied to Gaussian derivatives. First of all (following the second theorem) the minimal number of interpolation functions for the nth order Gaussian derivative is $n+1$. The interpolation functions themselves depend on the chosen θ_j. These angles of the basis filters should be chosen so that the response of the interpolation is robust with respect to noise. An appropriate choice is therefore to distribute the angles equally between $0°$ and $180°$. (Another choice for the θ_j may be motivated by the desire to obtain separable filters for x and y. See for details [Fre 91].)

Besides the well known interpolation functions for the first order Gaussian derivatives (see equation 3.25: $\theta_1 = 0°, \theta_2 = 90°, k_1(\phi) = \cos(\phi), k_2(\phi) = \sin(\phi)$) the following interpolation functions are obtained for the second order Gaussian derivatives (using $\theta_1 = 0°$, $\theta_2 = 60°$ and $\theta_3 = 120°$):

$$k_j(\phi) = \frac{1 + 2\cos(2(\phi - \theta_j))}{3} \text{ for } j = 1,2,3 \tag{3.29}$$

$$G^{\sigma}_{2,\phi} = k_1(\phi) G^{\sigma}_{2,0°} + k_2(\phi) G^{\sigma}_{2,60°} + k_3(\phi) G^{\sigma}_{2,120°} \tag{3.30}$$

3.1.2 Gabor filter

Gabor filters are local compact filters tuned to a spatial frequency band. Gabor filters are defined by modulating a Gaussian window with a cosine and an imaginary sine, giving an even and an odd filter pair [Gab 46]. The output of a Gabor filter may be represented as a complex number with a real and an imaginary part. This complex number can be expressed in polar coordinates as magnitude and direction.

A Gabor filter pair is compact in both space and frequency domains. In our experiments we use the two–dimensional formulation of the Gabor functions used by Daugman [Dau 93]:

$$g(x,y) = e^{-\pi\left(\frac{(x-x_0)^2}{\alpha^2}+\frac{(y-y_0)^2}{\beta^2}\right)} e^{-2\pi i(u_0(x-x_0)+v_0(y-y_0))} \tag{3.31}$$

where (x_0, y_0) are the central coordinates of the filter, (α, β) are the standard deviation used to determine the width and the length, and (u_0, v_0) specify the modulation in x and y directions, which has the spatial frequency $\omega_0 = \sqrt{u_0^2 + v_0^2}$ and direction $\theta_0 = \arctan(v_0/u_0)$. The Fourier domain transfer function $G(u, v)$ is given by:

$$G(u,v) = e^{-\pi\left((u-u_0)^2\alpha^2+(v-v_0)^2\beta^2\right)} e^{-2\pi i(x_0(u-u_0)+y_0(v-v_0))} \tag{3.32}$$

The main advantage of the Gabor filters is that one can freely choose the frequency (and therefore the scale) as well as the bandwidth of the filter.

For the design of a 1D Gabor filter, Westelius [Wes 92] proposes the tuning of the standard deviation α and the spatial frequency u_0. These two parameters determine the size and bandwidth of the filter. Given a certain spatial frequency u_0, Westelius chooses the radius of the frequency support so that the frequency function is sufficiently low at $u = 0$ (DC component). Therefore he defines the ratio between the DC component and the maximum value which should be smaller than a previously selected threshold P_{DC}. To generalize this ratio to 2D Gabor filters we set $\alpha = \beta$. The ratio is then given by (for $(x_0, y_0) = (0, 0)$):

$$\frac{G(0,0)}{G(u_0,v_0)} \leq P_{DC} \Rightarrow \alpha \geq \frac{\sqrt{-\ln P_{DC}}}{\sqrt{\pi}\omega_0} \tag{3.33}$$

Both the spatial support and the frequency support of an ideal Gabor function are infinite. To define a digital filter we must sample the function and limit its spatial support. The limit of the spatial size is chosen such that the amplitude of the filter is lower than a threshold P_{cut} for all positions outside the limit. The spatial radius $R = \sqrt{r_x^2 + r_y^2}$ of the filter is then:

$$\frac{\|g(r_x,r_y)\|}{\|g(0,0)\|} \leq P_{cut} \Rightarrow R \geq \alpha\frac{\sqrt{-\ln P_{cut}}}{\sqrt{\pi}} \tag{3.34}$$

One should notice that the chosen radius R affects the DC component of the signal. Therefore the DC component should be checked after truncation of the filter support.

By using Euler's equation ($e^{i\omega x} = \cos\omega x + i\sin\omega x$) we can rewrite equation 3.31 in polar form:

$$\begin{aligned} g(x,y) &= Re(g(x,y)) + Im(g(x,y)) && (3.35) \\ Re(g(x,y)) &= \cos\left(-2\pi\left(u_0(x-x_0)+v_0(y-y_0)\right)\right) e^{-\pi\left(\frac{(x-x_0)^2}{\alpha^2}+\frac{(y-y_0)^2}{\beta^2}\right)} && (3.36) \\ Im(g(x,y)) &= i\sin\left(-2\pi\left(u_0(x-x_0)+v_0(y-y_0)\right)\right) e^{-\pi\left(\frac{(x-x_0)^2}{\alpha^2}+\frac{(y-y_0)^2}{\beta^2}\right)} && (3.37) \end{aligned}$$

The real part $Re(g(x,y))$ can be interpreted as a second order derivative and the imaginary part $Im(g(x,y))$ can be interpreted as a first order derivative (both in direction of θ_0).

3.1.3 Color

Color is one of the most tempting and the most unused local descriptor for object recognition. It is tempting since for a large number of object classes (as for example flowers, trees, fruits, streets, faces) color can be used as the principal cue for identification and classification. Unfortunately color is relatively unstable to illumination changes (in particular to changes of illumination intensity and illumination color). Probably because of this color is rarely used for object recognition. In contrast to this widely accepted view, researchers have recently proposed color invariants which provide local characteristics invariant to illumination changes and illumination color shifts. In order to derive these invariants we need to introduce a general color model.

Usually the general color model is composed of three parts (with $\mathbf{x}$ a two dimensional image location and λ the wavelength of the light source):

- $E(\mathbf{x},\lambda)$ spectral energy distribution of the light source,
- $R(\mathbf{x},\lambda)$ reflectance function of the surfaces in the scene,
- $r_k(\lambda)$ the sensitivity function of the sensor k (often $k \in \{Red, Green, Blue\}$).

The signal $S_k(\mathbf{x})$ transmitted by the sensor k may be modeled by the following equation:

$$S_k(\mathbf{x}) = \int_\lambda E(\mathbf{x},\lambda)R(\mathbf{x},\lambda)r_k(\lambda)d\lambda \qquad (3.38)$$

Starting from (usually) three such sensor signals the goal is to derive a local descriptor dependent only on the reflectance function $R(\mathbf{x},\lambda)$ of the object. Such a local descriptor should be independent of the spectral energy distribution $E(\mathbf{x},\lambda)$ of the light source and should be independent of the sensor characteristics $r_k(\lambda)$.

The classical approach is to develop a "Color Constancy" algorithm which extracts (exactly or approximately) the reflectance function $R(\mathbf{x},\lambda)$ [Mal 85, Mal 86, For 90, Fun 91, Fin 95a, Fin 95b, Bar 96]. Since this is not possible for the general case many assumptions about the image, the object surface and the lighting conditions have to be made, which limits the application of such algorithms to constrained or controlled circumstances.

A relatively small number of authors propose to calculate color invariants which depend mainly on the reflectance function $R(\mathbf{x},\lambda)$. Such invariants have been proposed for example by Nagao [Nag 95] and Funt and Finlayson [Fun 95].

Narrow band sensor model

In order to simplify the general color model, several authors use the narrow band sensor model:

$$r_k(\lambda) = \delta(\lambda - \lambda_k) \tag{3.39}$$

with λ_k the wavelength of maximal sensitivity for the sensor k. The general color model becomes a simple multiplication:

$$S_k(\mathbf{x}) = R(\mathbf{x}, \lambda_k)E(\mathbf{x}, \lambda_k) \tag{3.40}$$

Even though this assumption is certainly an oversimplification, it allows the calculation of several invariants shown to be robust to illumination intensity changes [Col 96]. We will introduce the invariants proposed by Nagao and by Funt and Finlayson.

Color Invariant proposed by Nagao

Under the assumption of $E(\mathbf{x}, \lambda_i) = E(\lambda_i)$ (constant illumination) over the entire image we can calculate the relation of different sensors i and j (at the same image location $\mathbf{x}$):

$$\gamma_{ij} = \frac{S_i(\mathbf{x})}{S_j(\mathbf{x})} = \frac{R(\mathbf{x}, \lambda_i)E(\lambda_i)}{R(\mathbf{x}, \lambda_j)E(\lambda_j)} \tag{3.41}$$

The same relation γ'_{ij} can be calculated in another image I' (of the same object at a different illumination) at the corresponding location $\mathbf{x}'$. The relation between γ_{ij} and γ'_{ij} is constant over the entire image:

$$\frac{\gamma_{ij}}{\gamma'_{ij}} = \epsilon_{ij} = \frac{E(\lambda_i)}{E(\lambda_j)} / \frac{E'(\lambda_i)}{E'(\lambda_j)} \tag{3.42}$$

Since this constant depends only on the energy distribution of the light source it is possible to normalize the image with this constant in order to obtain an illumination independent color image.

Color Invariant proposed by Funt and Finlayson

Funt and Finlayson propose to calculate the response relation for two close points $\mathbf{x}$ and $\mathbf{y}$ of the same sensor k:

$$\frac{S_k(\mathbf{x}, \lambda_k)}{S_k(\mathbf{y}, \lambda_k)} = \frac{R(\mathbf{x}, \lambda_k)E(\mathbf{x}, \lambda_k)}{R(\mathbf{y}, \lambda_k)E(\mathbf{y}, \lambda_k)} \tag{3.43}$$

Under the assumption that $E(\mathbf{x}, \lambda_k) = E(\mathbf{y}, \lambda_k)$, corresponding to a locally constant energy distribution, we obtain the following invariant:

$$\frac{S_k(\mathbf{x},\lambda_k)}{S_k(\mathbf{y},\lambda_k)} = \frac{R(\mathbf{x},\lambda_k)}{R(\mathbf{y},\lambda_k)} \tag{3.44}$$

In order to avoid the division operation Funt and Finlayson propose to calculate differences of logarithms:

$$\log S_k(\mathbf{x},\lambda_k) - \log S_k(\mathbf{y},\lambda_k) = \log R(\mathbf{x},\lambda_k) - R(\mathbf{y},\lambda_k) \tag{3.45}$$

We conclude from experimental results [Col 96] that in particular the latter invariants (equation 3.45) are stable over a large range of illumination intensity changes. Even in the presence of illumination color shifts the invariants behave well. In the presence of significant illumination color shifts the invariants did not perform well probably due to the oversimplification of the narrow band sensor model.

3.2 Normalization of filter output

The effects of variation in signal intensity can be eliminated by normalizing the inner product of a filter with the signal during convolution. Normalization should be considered from at least two points of view. The first point concerns how well the normalized convolution behaves in the presence of additive noise. The second point concerns how the normalized convolution responds to variations in signal intensity due to differences in ambient illumination intensity, aperture setting or digitizer gain. In this section we will introduce several different normalization techniques.

In section 3.2.4 the robustness of the normalization techniques relative to additive Gaussian noise is examined. Section 5.3 describes experiments where good robustness to illumination intensity changes is achieved by using energy or variance normalization.

3.2.1 Normalization by mean and variance

By variance (or variance–mean) normalization we refer to subtracting the mean of each neighborhood and then dividing by the variance of the neighborhood. Variance normalization is in some special cases invariant to scale for example in the case of binary images. But this is not the case for more arbitrary gray level signals. In these cases the equivariance property of Gabor filters and Gaussian derivatives makes the explicit choice of the scale possible. Scale then becomes a free parameter which we can tune in order to detect an object (see sections 3.1.1 and 3.1.2). The variance normalization is defined by:

$$I_{var}(x,y) = \frac{\sum_{i,j=-m,-n}^{m,n}(I(x+i,y+j) - \overline{I(x,y)})M(i,j)}{\sqrt{\sum_{i,j=-m,-n}^{m,n}(I(x+i,y+j) - \overline{I(x,y)})^2}\sqrt{\sum_{i,j=-m,-n}^{m,n} M(i,j)^2}} \tag{3.46}$$

with

$$\overline{I(x,y)} = \frac{1}{(2m+1)(2n+1)} \sum_{i,j=-m,-n}^{m,n} I(x+i,y+j) \tag{3.47}$$

The mean of the neighborhood could be replaced by the mean over the whole image, making the assumption that the illumination intensity changes are the same in the entire image. Since this cannot be assumed in general it is a rather unrealistic assumption but computationally less expensive.

Variance normalization is relatively sensitive to additive Gaussian noise. This can be seen by considering the effects of variance normalization on a constant signal. Subtracting the mean brings the signal to zero, and its variance is then zero. Digitizing an image necessarily results in some additive Gaussian noise due to quantization and sampling. Even when the noise energy is miniscule, the normalized filter output will respond purely to the noise. This makes variance–normalization unusable in our context.

3.2.2 Normalization by energy

The energy of a signal is the sum of the square of the coefficients. Energy normalization sets the energy of the filter coefficients, and the energy of each neighborhood to unity. The cost of normalizing the individual neighborhood energies can be minimized by incrementally computing a running sum of squares of the neighborhood during convolution. Dividing by neighborhood energy removes variations in signal strength which may be due to light source intensity variation, and thus provide a filter output vector which is invariant to illumination intensity (see experiments in section 5.3):

$$I_{ene}(x,y) = \frac{\sum_{i,j=-m,-n}^{m,n} I(x+i,y+j)M(i,j)}{\sqrt{\sum_{i,j=-m,-n}^{m,n} I(x+i,y+j)^2}\sqrt{\sum_{i,j=-m,-n}^{m,n} M(i,j)^2}} \tag{3.48}$$

3.2.3 Normalization by max–min

The max–min normalization makes all filter outputs invariant to illumination intensity changes, provided that input is within the linear region of the dynamic range of the CCD–camera [Bob 95].

As one would expect, the max–min–normalization is extremely sensitive to noise. Especially in the case of nearly binary images.

$$I_{maxmin}(x,y) = \frac{\sum_{i,j=-m,-n}^{m,n} I(x+i,y+j)M(i,j)}{\max_{i,j} I(x+i,y+j) - \min_{i,j} I(x+i,y+j)} \tag{3.49}$$

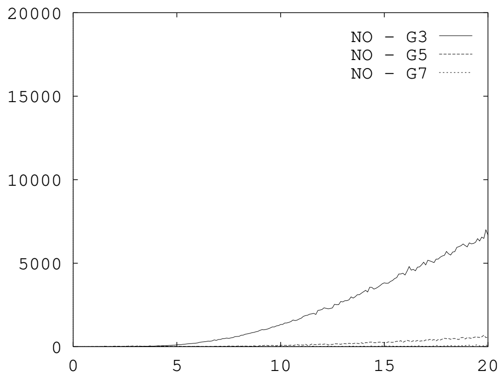
3.2.4 The robustness of normalization techniques in the presence of additive Gaussian noise

This section reports results of an experiment which was designed to determine the sensitivity of the different normalization techniques with respect to additive Gaussian noise. For this experiment we used 8 artificial images. We will summarize the results of two images, which we call *Sine* and *Grid* image. The *Sine* image contains a sine–curve with a wavelength of 45 pixels. The Grid image is basically a binary image with a Grid of 30×30 squares. The two images are shown in figure 3.1.

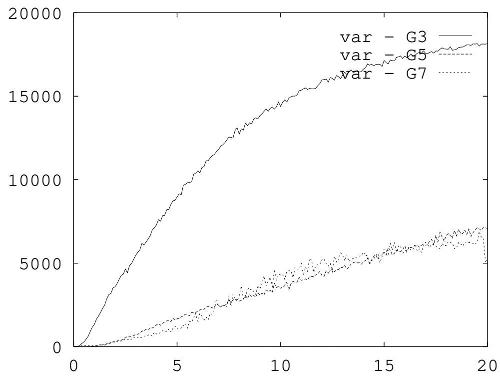
Figure 3.2 shows results for the *Sine* image. To the *Sine*–image we added Gaussian noise with standard deviation $\sigma = 1, 2, 3, \ldots, 20$ (abscissa in the diagrams). We store the two–dimensional histogram of the initial image (which is equivalent to $\sigma = 0$). This histogram is then compared

Figure3.1: The Sine–image and the Grid image

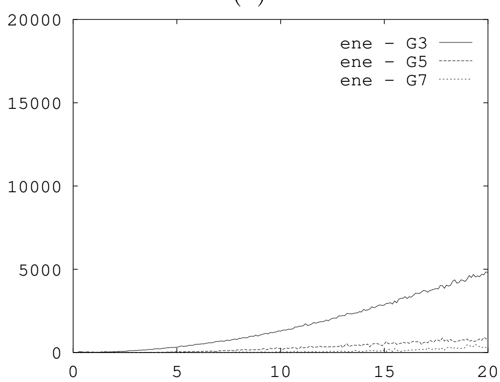
(by using χ^2_{qv} as similarity measure, see section 5.1.3) to the histograms with additive Gaussian noise. This comparison measurements corresponds to the ordinate of the diagrams.



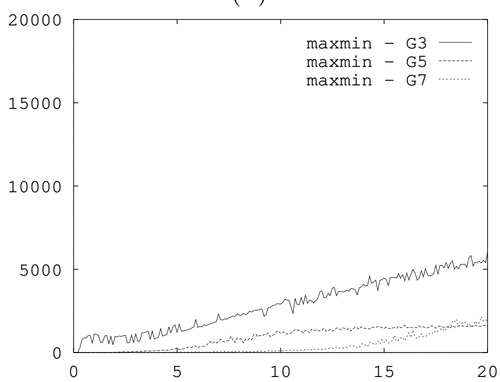
(a)



(b)



(c)



(d)

Figure3.2: Sine: (a) Robustness with no normalization. (b) Robustness with variance normalization. (c) Robustness with energy normalization. (d) Robustness with max–min normalization

In these diagrams we are mainly interested in the relative behavior of the different filter normalization techniques, rather than the absolute value of the χ^2_{qv} statistics between the images (which depends strongly on the design parameters of the histograms). In this experiment we used three Gabor filters (see section 3.1.2):

- *G3*: Gabor filter with wavelength of 2.8 pixels (7×7 window) in x and in y direction,
- *G5*: Gabor filter with wavelength of 5.7 pixels (15×15 window) in x and in y direction,
- *G7*: Gabor filter with wavelength of 11.3 pixels (30×30 window) in x and in y direction.

The first statement we can make is, that the Gabor filters are relatively robust to additive Gaussian noise (figure 3.2(a)). This can be explained by the fact that one part of the Gabor function is a Gaussian smoothing function. The stability of the Gabor filter increases with the size of the Gabor filter. Only in the case of variance normalization do the Gabor filters not behave well (see figure 3.2(b)). The second statement compares the different normalization techniques: *no* normalization behaves rather well (figure 3.2(a)). The *variance* normalization on the other hand disturbs the nice behavior of the Gabor filters (figure 3.2). The *max–min* normalization (figure 3.2(d)) gives more unstable even though still good results. The best results are obtained by the *energy* normalization (figure 3.2(c)).

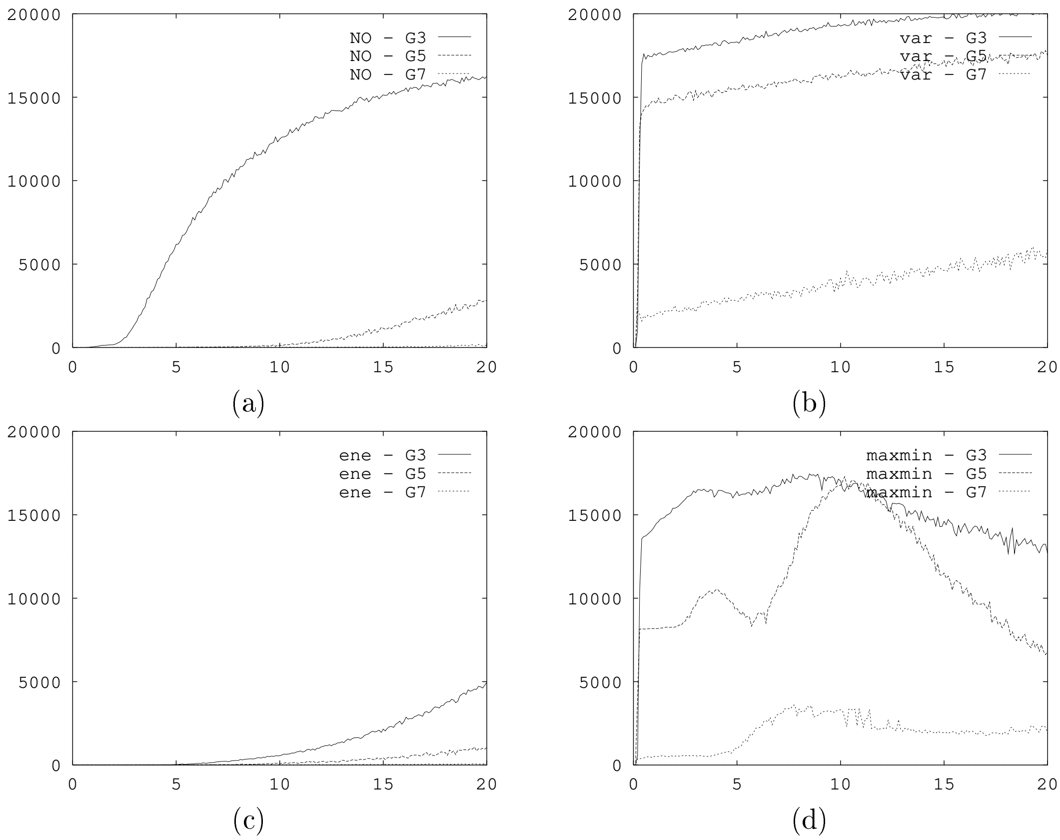


Figure3.3: Grid: (a) Robustness with no normalization. (b) Robustness with variance normalization. (c) Robustness with energy normalization. (d) Robustness with max–min normalization

Figure 3.3 shows the robustness of the different normalization techniques for the *Grid* image. In this case *no* normalization behaves much worse than in the case of the *Sine* image. Both the *max–min* and the *variance* normalization do not provide satisfactory results. In particular the variance normalization amplifies the influence of the noise. The best results are obtained by the *energy* normalization which are stable.

In the following chapters we will use only *energy* normalization since it seems to be the most robust normalization for the considered filters with respect to additive noise. The following chapter shows satisfactory results with this normalization in the recognition experiments.

3.3 Conclusion

The chapter described local descriptor based on Gaussian derivatives, Gabor filters and color information. Statistics of vectors of such local descriptors will be used for the representation of objects. Typically we use Gaussian derivatives, introduced in section 3.1.1. The steerability of Gaussian derivatives with respect to image plane rotation and the equivariance property to scale changes have been introduced. In particular, these properties make Gaussian derivatives appropriate for the representation of objects in our context. Gabor filters posse the same properties, but Gaussian derivatives can be computed more rapidly due to recursive implementation [Der 93]. The described color invariance offers the possibility of robust local descriptor in the presence of changes in the intensity and color of the light source.

Section 3.2 introduced normalization techniques of filter responses, which may provide robustness of the filter with respect to noise and illumination intensity changes. From an analysis of the robustness of Gabor filter with respect to additive Gaussian noise we concluded that energy normalization provides the best results. Because of this result and the experiments described in section 5.3 (robustness with respect to illumination intensity changes), the experiments described in the reminder of the thesis use energy normalization for filter responses.

As generalization of the color histogram approach, the next chapter develops the statistical representation of objects. The statistical representation is based on an arbitrary set of vectors of local characteristics called measurements. Such vectors may be given by a combination of local descriptor introduced here. In general, we can imagine any combination, which may be appropriate for the representation of a particular object set.

Chapter 4

Statistical object representation

The chapter introduces and motivates the use of multidimensional receptive field histograms for statistical representation of the appearance of objects (section 4.1). The main idea is to represent 3D objects by the probability density function of 2D local characteristics, 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. In this work we use local descriptors introduced in chapter 3. But the same method can (and should) be applied to other local descriptors (for example grey value moments or geometric invariants).

Section 4.2 develops an analogy between object recognition and information theory. An interesting result of this analogy is the application of transinformation of the statistical object representation as a quantitative measurement for the evaluation of the used local descriptors (with respect to the considered objects).

4.1 Statistical object representation

The following develops a statistical object representation based on the probability density function of local characteristics of an object o_n. Let's consider we have chosen a fixed measurement set M of local characteristics m_k. 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. Such changes should be modeled within the probability density function. Four categories of possible changes can be listed (see figure 4.1):

Similarity transformation in the image plane: 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 4.1).

3D transformation of the object: two rotational degrees of freedom (r_x and r_y) exist in addition to the similarity transformation (see figure 4.1).

Scene changes include partial occlusion and background change.

Imaging conditions change accordingly to light changes and different types of signal disturbance as signal noise, quantization error and blur. The principal characteristic of these changes is that they cannot be controlled and/or predicted.

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

$$p(M|o_n, R, T, S, I) \tag{4.50}$$

where M is the set of local characteristics 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 and I the imaging conditions.

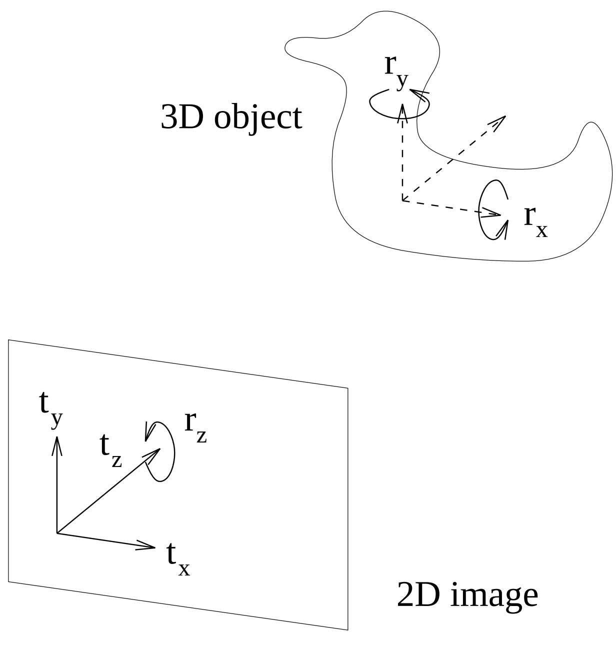


Figure4.1: Different components of rotation and translation of a 3D object

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 [Int 93]. The most effective way to reduce the number of free parameters is to choose local characteristics which are invariant in relation to different parameters. Such invariant features are used by many researchers [Bur 92, Mun 92, Mun 93] and applied successfully in various ways. Unfortunately the obtained invariants are very restrictive to certain types of objects. As motivated in chapter 3 we do not want to restrict the applicability of our approach by the use of a particular type of invariants.

Another way to reduce the number of free parameters of the probability density function (equation 4.50) is to use local characteristics which are robust with respect to certain changes. Robust means that the local characteristics are changing slowly with the considered changes. Typically a certain range of change can be identified where the local characteristics are nearly constant. Under consideration of a known range of change, robust local characteristics become nearly-invariant descriptors. We are convinced that many local characteristics can be calculated in a robust manner without being invariant in general.

One of the above mentioned category of changes, the *imaging conditions*, is characterized by changes which cannot be controlled. 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 of the used local characteristics to such changes therefore demands special consideration. Section 3.2.4 examines the robustness of local characteristics and different normalization techniques in the presence of additive Gaussian noise. A high robustness to light intensity changes can be concluded from experiments described in section 5.3.

Scene changes like *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 [Hor 95] 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 chapter 7 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.

A complete 3D transformation of an object is given by three parameter for a *translation* and another three parameter for an arbitrary *rotation* of the object. The three translational parameter and the rotational parameter r_z – which models the image plane rotation, see figure 4.1 – are compromised in the similarity transformation. The remaining two rotational parameters r_x and r_y represent a viewpoint change.

A *similarity transformation* can be managed by transformation of the observed image pattern. This includes the translation of the image pattern parallel to the image plane (t_x and t_y), rotation inside the image plane (r_z) and scale change of the image pattern (t_z). All four parameters can be treated by a the transformation of the image pattern itself, including interpolation and sub–sampling. The precision of the similarity transformation is limited by the resolution of the image and the size of the image pattern. The higher the resolution of the image the more scale change t_z can be managed. And the larger the image pattern itself the smaller the expected error of the transformation becomes.

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 dimensionality of the probability density function is reduced by two dimensions. And thirdly the estimation problem of the probability density function 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.

As mentioned above, the third translational parameter t_z can be treated directly by the transformation of the image pattern. Throughout the thesis 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 proposed for example by Schmid [Sch 96h]. 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 changes* of the observer. Several authors [Bur 90, Cle 91] show that the non–existence of viewpoint invariant descriptors for the general case. Nevertheless useful descriptors exist in special cases [Mun 92, Mun 93]. 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 4.50) are three components of the rotation and one component of the translation:

$$p(M|o_n, r_x, r_y, r_z, t_z) \tag{4.51}$$

By considering an L–dimensional vector m_k of local characteristics 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.

4.1.1 Estimation and representation by multidimensional receptive field histograms

Different possibilities exist in order to estimate and represent the probability density function (equation 4.51) 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 [Hor 95] 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 characteristics.

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 [Pop 94]. The main advantage of histogramming is that the training samples are well represented. This property is desirable in our context since all information and in particular the discriminant information is preserved. On the other hand kernel functions typically allow the generalization from training samples. In the context of histogramming the generalization capability can be obtained by a sufficient number of training samples. Sufficient means that the order of training samples should be at least in the order of the cells of the histogram. By reducing the number of histogram cells for example we eventually obtain a generalization capability for the sake of discriminant information. In our context and since we do not represent image plane translations (t_x and t_y) the number of training samples is directly given by the size of the object in the image. As a result the order of training samples – respectively image measurements – is in the same order as the size of the object. Typically the number of different measurements for an object is much smaller than the number of training samples. We can therefore expect that histograms generalize from the training samples.

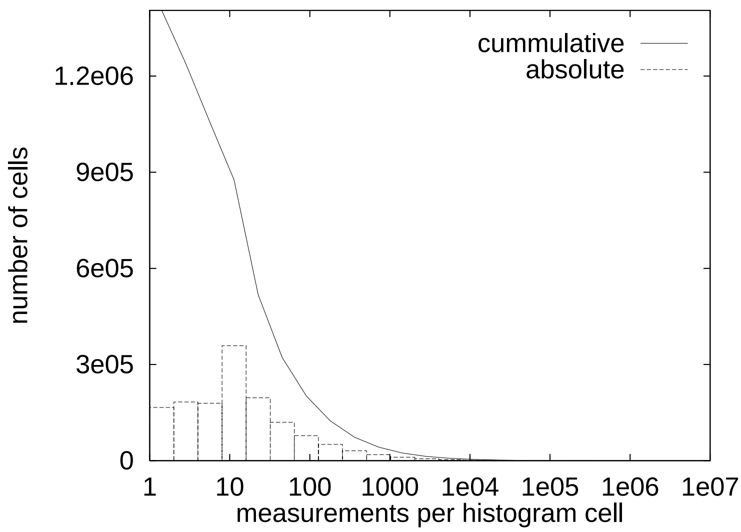


Figure4.2: Distribution of measurements per histogram cells. The horizontal axis shows the number of measurements per histogram cell whereas the vertical axis shows the corresponding number of histogram cells

In the following we describe the order of histogram cells in the experiment described in section 7.3. In this experiment we used the largest histograms described in the thesis: six–dimensional histograms with a resolution of 32. The order of all possible histogram cells is therefore 10^9. Due to their important size the histograms are particularly difficult to estimate. As a matter of fact not all 10^9 histogram cells – respectively the corresponding six–dimensional measurement vectors – occur in the images of the used 103 objects. In average only 5000 histogram cells occurs at least one time in the image. Each of the histograms corresponds to a particular viewpoint, image plane rotation and scale.

It is interesting to examine the overall number of measurements which occur at least one time in any of the histograms. Figure 4.2 shows the distribution of measurements over the histogram cells of all 6367 calculated histograms. The actual order of histogram cells which occur at least one time is about 10^6. The order of histogram cells which occur at least 100 times is about 10^5. Using a 512×512 image the number of measurements per image is also in the order 10^5. In conclusion we can expect histograms to generalize from the training samples. We'd like to point out that this example corresponds to the most difficult case encountered during the thesis. The generalization capability can be improved either by reducing the dimensionality of the histograms or by reduction of the resolution per histogram axis.

Consequently we represent the probability density function of a certain object by several multidimensional histograms over the measurement set M. 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:

$$H(M|o_n, r_x, r_y, r_z, t_z) \tag{4.52}$$

In order to obtain these histograms we have to take several images of the object. The number of images can be reduced considerably by using the steerability to image plane rotation and the equivariance property of local characteristics to scale changes. The steerability and equivariance property of Gaussian derivatives is described in section 3.1.1. 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). Section 6.5 examines the number of histograms which are needed for the representation of a 3D object. We conclude from the experiments that, because of the robustness of multidimensional receptive field histograms to viewpoint changes, already a small number of histograms are sufficient in order to obtain high recognition rates.

Throughout the thesis we only use multidimensional histograms for the representation (or better estimation) of the probability density function. It is worth mentioning that this is not the most efficient way for the 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 thesis is to show that the representation of objects by the probability density function of their local characteristics 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 (chapters 5 and 6) and probabilistic object recognition (chapter 7).

4.2 The application of information theory to object representation

In the following we express the object recognition process in terms of the transmission of information through a (noisy) channel. This analogy is most appropriate for a statistical object representation as introduced in the previous section. But the analogy can be applied to a wide variety of object recognition processes. The only assumption for the validity of the analogy is that the "messages" (or measurements) m_k obtained from the object are member of a certain finite measurement set M.

The following section 4.2.1 summarizes basic concepts from information theory (see for example [Rez 95]) and gives a first interpretation of the concepts in the context of object recognition. The reader familiar with information theory should move directly to section 4.2.2 which discusses in more detail the interpretation of different entropies in the context of object recognition. Section 4.2.3 introduces transinformation (or mutual information) of the object recognition process. In particular transinformation may be applied for the evaluation of an employed measurement set M. Experiments show that we can predict the relative performance of an object recognition process using different measurement sets M. Section 4.2.4 finally defines the information theoretical concepts of capacity, redundancy and efficiency in the context of object recognition.

4.2.1 Information measurement

Let's assume a partition of the sample space Ω_X into a finite number of mutually exclusive events x_n, whose probabilities $p(x_n)$ are known. The events x_n are a complete partition in the sense that:

$$\bigcup_{n=1}^{N} x_n = \Omega_X \tag{4.53}$$

$$\sum_{n=1}^{N} p(x_n) = 1 \tag{4.54}$$

A probability scheme with these properties is called a *complete finite probability scheme*. The fundamental problem of interest in information theory is to define a measure of *uncertainty* for such a probability scheme. Shannon and Wiener have suggested to use the following well known equation:

$$H(X) = -\sum_{n=1}^{N} p(x_n) \log(p(x_n)) \tag{4.55}$$

Using the quantity $I(x_n) = -\log(p(x_n))$ as the measure of *self–information* of the event x_n we can interpret $H(X)$ as the average self-information of each event x_n:

$$H(X) = \overline{I(x_n)} = \sum_{n=1}^{N} p(x_n) I(x_n) \tag{4.56}$$

An object set $O = \Omega_O = \bigcup_{n=1}^{N} o_n$ with the probabilities $p(o_n)$ also form a complete finite probability scheme as defined above. Therefore we can follow equation 4.55 in order to define the average self–information $H(O)$ of each object o_n:

$$H(O) = -\sum_{n=1}^{N} p(o_n)\log(p(o_n)) \tag{4.57}$$

The same analogy can be applied to a measurement set $M = \Omega_M = \bigcup_{k=1}^{K} m_k$ which also forms a complete finite probability scheme:

$$H(M) = -\sum_{k=1}^{K} p(m_k)\log(p(m_k)) \tag{4.58}$$

In the context of transmission of information through a channel the relation between the input symbols and the output symbols needs to be known. The following therefore develops an information measurement for the two–dimensional case.

Information measurement for the two–dimensional case:

The measurement $H(X)$ of *uncertainty* or *information* can be generalized to a two-dimensional complete finite probability scheme. Such a probability scheme is given by two sample spaces Ω_X and Ω_Y where we select complete event sets x_n and y_k in the sense of equation 4.53 and 4.54. Each event x_n of Ω_X may occur in conjunction with any event y_k of Ω_Y. Therefore the product space $\Omega_X \times \Omega_Y$ forms a complete set of events with the following probability matrix:

$$P(X \wedge Y) = \begin{pmatrix} p(x_1 \wedge y_1) & p(x_1 \wedge y_2) & \dots & p(x_1 \wedge y_K) \\ p(x_2 \wedge y_1) & p(x_2 \wedge y_2) & \dots & p(x_2 \wedge y_K) \\ \vdots & & \ddots & \vdots \\ p(x_N \wedge y_1) & p(x_N \wedge y_2) & \dots & p(x_N \wedge y_K) \end{pmatrix} \tag{4.59}$$

Therefore we have three complete probability schemes, namely $P(X)$, $P(Y)$ and $P(X \wedge Y)$. We can define three corresponding entropies:

$$H(X) = -\sum_{n=1}^{N} p(x_n)\log(p(x_n)) \tag{4.60}$$

$$H(Y) = -\sum_{k=1}^{K} p(y_k)\log(p(y_k)) \tag{4.61}$$

$$H(X \wedge Y) = -\sum_{n=1}^{N}\sum_{k=1}^{K} p(x_n \wedge y_k)\log(p(x_n \wedge y_k)) \tag{4.62}$$

with

$$
\begin{aligned}
p(x_n) &= \sum_{k=1}^{K} p(x_n \wedge y_k) && (4.63)\\
p(y_k) &= \sum_{n=1}^{N} p(x_n \wedge y_k) && (4.64)
\end{aligned}
$$

$H(X)$ is called marginal entropy of X, $H(Y)$ marginal entropy of Y and $H(X \wedge Y)$ joint entropy. We can define two further entropies: the conditional entropies $H(X|Y)$ and $H(Y|X)$. In order to obtain the equation for $H(X|Y)$ we start with the self-information of the event $(x_n|y_k)$ which is:

$$
I(x_n|y_k) = -\log(p(x_n|y_k)) \qquad (4.65)
$$

The average Information $H(X|y_k)$ given a certain y_k is then:

$$
H(X|y_k) = \overline{I(x_n|y_k)} = -\sum_{n=1}^{N} p(x_n|y_k)\log(p(x_n|y_k)) \qquad (4.66)
$$

The average conditional entropy $H(X|Y)$ is therefore calculated by:

$$
\begin{aligned}
H(X|Y) &= \overline{H(X|y_k)} = \sum_{k=1}^{K} p(y_k) H(X|y_k) && (4.67)\\
&= -\sum_{k=1}^{K} p(y_k) \sum_{n=1}^{N} p(x_n|y_k)\log(p(x_n|y_k)) && (4.68)\\
&= -\sum_{k=1}^{K}\sum_{n=1}^{N} p(x_n \wedge y_k)\log(p(x_n|y_k)) && (4.69)
\end{aligned}
$$

Analogous we can derive the equation for the conditional entropy $H(Y|X)$:

$$
H(Y|X) = -\sum_{k=1}^{K}\sum_{n=1}^{N} p(x_n \wedge y_k)\log(p(y_y|x_n)) \qquad (4.70)
$$

4.2.2 Application of information theory to object recognition

In the previous section we derived five entropies for a two-dimensional complete finite probability scheme without giving their interpretation. In information theory a two-dimensional probability scheme typically describes a communication network: x_n are the possible inputs (or symbols of the input alphabet) and y_k are the possible outputs of the network. Each input x_n is "transformed" by the communication channel to possible outputs y_k, whereas the joint probability matrix (equation 4.59) describes the characteristics of the channel.

In the context of object recognition the possible "inputs" are the different objects o_n and the possible "outputs" are the measurements or symbols m_k which can be extracted from the image of an object. The channel corresponds to the transformation of the objects to the measurement space. Therefore, the communication channel corresponds to the recognition process as a whole. The characteristics of the recognition process are given by the joint probability matrix (equation 4.59). Using this analogy between a communication network and the object recognition process we can interpret the five entropies $H(O)$, $H(M)$, $H(O \wedge M)$, $H(O|M)$ and $H(M|O)$ in the following way:

- $H(O)$ (equation 4.60) is the average information of each object o_n,
- $H(M)$ (equation 4.61) is the average information of each measurement m_k,
- $H(O \wedge M)$ (equation 4.62) is the overall uncertainty of the recognition process,
- $H(M|O)$ (equation 4.70) gives an indication of the average "noise" or error of the recognition process,
- $H(O|M)$ (equation 4.69) indicates the overall quality of the recognition process. The smaller $H(O|M)$ the better the object set O can be recognized with the measurement set M.

Since we consider that all objects o_n do have an a priori known and fixed probability $p(o_n)$, the entropy $H(O)$ is constant. In the reminder of the thesis we typically assume that all objects are equally probable ($p(o_n) = 1/N$). The entropy $H(O)$ becomes then: $H(O) = \log(N)$.

The remaining four entropies will change significantly, when the measurement set M is changed. Perhaps the most interesting entropy is $H(O|M)$ which indicates the error of the recognition process. In order to show the influence of the chosen measurement set M we develop $H(O|M)$ (equation 4.69):

$$\begin{aligned}
H(O|M) &= -\sum_{k=1}^{K}\sum_{n=1}^{N} p(o_n \wedge m_k)\log(p(o_n|m_k)) \\
&= -\sum_{k=1}^{K}\sum_{n=1}^{N} p(o_n \wedge m_k)\log\frac{p(o_n \wedge m_k)}{p(m_k)} \\
&= -\sum_{k=1}^{K}\sum_{n=1}^{N} p(o_n \wedge m_k)\log(p(o_n \wedge m_k)) + \sum_{k=1}^{K}\sum_{n=1}^{N} p(o_n \wedge m_k)\log(p(m_k)) \\
&= H(O \wedge M) + \sum_{k=1}^{K} p(m_k)\log(p(m_k)) \\
&= H(O \wedge M) - H(M) \qquad (4.71)
\end{aligned}$$

By minimizing $H(O|M)$, the average error of the recognition process is reduced. Following equation 4.71 this can be obtained by minimizing the overall uncertainty $H(O \wedge M)$ and by maximizing $H(M)$. $H(M)$ has its maximum when all measurements m_k are equally-probable. $H(O \wedge M)$ is minimal, when for each object o_n there exist exactly one measurement m_k so that $p(o_n|m_k) = 1$ and $p(m_k|o_n) = 1$ (this corresponds to the case that each row of matrix 4.59 contains at exactly one position the value $1/N$). In other words, the smaller $H(O \wedge M)$

the more significant are the measurements m_k in average. Therefore equation 4.71 provides a possibility to compare numerically different measurement sets m_k for the same object set o_n. Equation 4.71 can be applied in the context of many object recognition processes, as soon the calculation or estimation of the joint probability matrix 4.59 is possible.

4.2.3 Transinformation of the object recognition process

In information theory the mutual information contained in the event pair (x_n, y_k) is the basis of the calculation of the *transinformation* of the channel. By applying the analogy between the communication network and the object recognition process the calculation of the transinformation of an object/measurement pair (o_n, m_k) becomes:

$$T(o_n, m_k) = \log \frac{p(o_n \wedge m_k)}{p(o_n)p(m_k)} \tag{4.72}$$

Therefore the average transinformation per symbol pair is:

$$T(O, M) = \overline{T(o_n, m_k)} = \sum_{n=1}^{N} \sum_{k=1}^{K} p(o_n \wedge m_k) \log \frac{p(o_n \wedge m_k)}{p(o_n)p(m_k)} \tag{4.73}$$

This entropy indicates a measure of the information transmitted through the channel (= recognition process). For this reason it is known as *transinformation* of the channel. By applying the definition (equation 4.73) we obtain:

$$\begin{aligned} T(O, M) &= H(O) + H(M) - H(O \wedge M) && (4.74) \\ &= H(O) - H(O|M) && (4.75) \\ &= H(M) - H(M|O) && (4.76) \end{aligned}$$

Following equation 4.75, the information transmitted by the channel (respectively by the recognition process) can be maximized by the minimization of $H(O|M)$ (assuming $H(O)$ to be constant as mentioned above). Or by using equation 4.76 we can say, that the maximization of $H(M)$ together with the minimization of $H(M|O)$ also result in the maximization of the transinformation. Similar results have been obtained from the analysis of equation 4.71.

Using the fact that $H(O \wedge M) \geq \max(H(O), H(M))$ in equation 4.74 we obtain the following bound of the transinformation $T(O, M)$:

$$T(O, M) \leq \min(H(O), H(M)) \tag{4.77}$$

Since we consider $H(O)$ to be constant (all objects o_n are equally probable), the upper bound of the transinformation $T(O, M)$ can be increased by $H(M)$ until $H(M)$ equals $H(O)$.

The most interesting idea of using transinformation is the possibility to compare different measurement sets M for a certain object set O. By applying the transinformation only onto a subset of O we obtain the most appropriate measurement set M for the discrimination of objects of that subset.

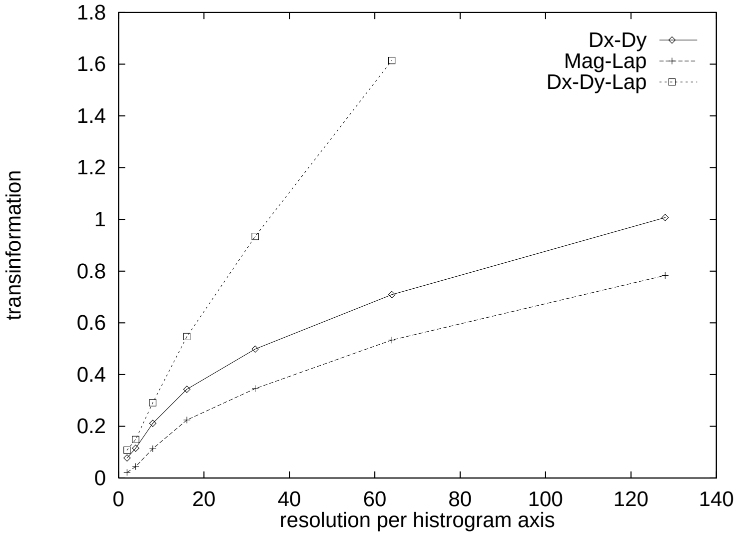


Figure4.3: The transinformation for 100 objects and for different measurement sets: Dx-Dy and Mag-Lap correspond to two-dimensional histograms and Dx-Dy-Lap corresponds to a three dimensional histogram. The horizontal axis shows different resolutions for the histogram axis and the vertical axis show the transinformation for a particular resolution

In order to illustrate the application of the transinformation for the evaluation of different measurement sets we have calculated the transinformation for 100 objects as a function of different filter-combinations at different resolutions. Here we used the Columbia image database with 100 objects which is shown in figure A.6 of the appendix. Figure 4.3 shows results of this evaluation. The three graphs correspond to three different measurement combinations: *Dx-Dy* corresponds the first Gaussian derivative in $x-$ and in $y-$direction. *Mag-Lap* is corresponds also to a two–dimensional measurement vector, namely the magnitude of the first Gaussian derivative and the Laplacian operator. *Dx-Dy-Lap* finally corresponds to a three-dimensional measurement vector of the first derivatives in $x-$ and $y-$direction and the Laplacian operator (we use the definition of Gaussian derivatives which we described in section 3.1.1). For each of the filter combinations we calculated histograms with different resolution, namely with resolution of 2, 4, 8, 16, 32, 64 and 128 per histogram axis.

Figure 4.3 shows that the choice of the measurement set plays an important role for the transinformation of the recognition process. The graph shows a significant increase of the transinformation for the three-dimensional measurement set *Dx-Dy-Lap* relative to the two-dimensional measurement sets *Dx-Dy* and *Mag-Lap*. This can be explained by the fact that *Dx-Dy-Lap* contains one more independent dimension. This increase is expected to be even more important by adding more independent dimensions. It is interesting to note that the result of *Dx-Dy* and *Mag-Lap* are qualitatively similar. Nevertheless *Dx-Dy* gives better results than *Mag-Lap*. This is due to the fact that *Dx-Dy* preserves the rotational information of the first derivative whereas *Mag-Lap* is rotation invariant.

Figure 4.4 shows the relation between the transinformation (figure 4.3) and the recognition rate for the different measurement sets. We used the same measurement sets and the same database images as in figure 4.3. As test images we used 100 images with a viewpoint change of

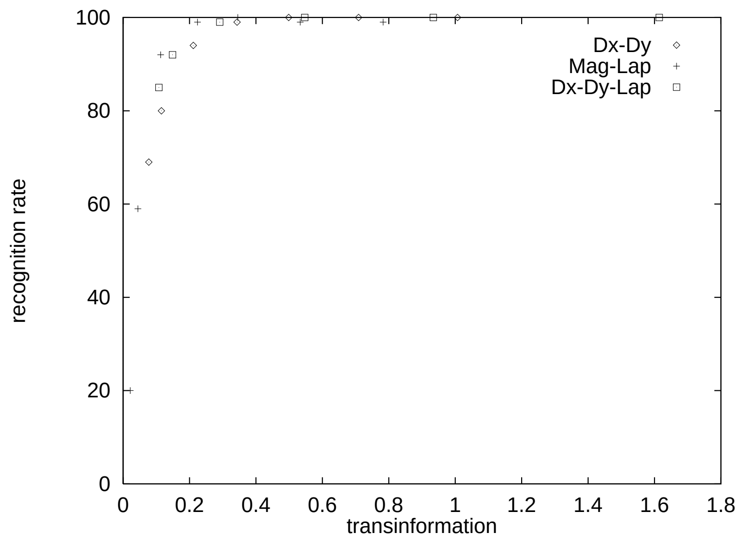


Figure4.4: The relation between the transinformation for 100 objects (for different measurement sets, see figure 4.3) and the recognition rate of 100 test images. We observe that all points lie on approximately the same curve, which shows the strong relation between the transinformation and the recognition rate. Therefore the transinformation can be used for the prediction of the recognition rate

5° relative to the database images. The figure shows the strong relation between the transinformation and the recognition rate. We are therefore convinced that we can predict the quality of recognition based on the transinformation of the object recognition process.

4.2.4 Capacity, redundancy and efficiency of the object recognition process

For completeness we will apply the information theoretical concepts of capacity, redundancy and efficiency to the proposed analogy to object recognition.

In information theory the channel capacity is defined as the maximum of the transinformation:

$$Capacity = \max_{P(O)} T(O, M) = \max_{P(O)} (H(M) - H(M|O)) \tag{4.78}$$

The maximization is with respect to all possible sets of probabilities $P(O)$ of the input, that means of the objects o_n. Therefore the capacity is calculated for a specific probability scheme $p(m_k|o_n)$.

Absolute redundancy is defined as the difference of the $Capacity$ and the transinformation:

$$Redundancy_{absolut} = Capacity - T(O, M) \tag{4.79}$$

Relative redundancy is defined as the relation of $Redundancy_{absolut}$ to $Capacity$:

$$Redundancy_{relative} = \frac{Capacity - T(O, M)}{Capacity} \tag{4.80}$$

Efficiency finally is defined as $Effiency = 1 - Redundancy_{relative}$.

4.3 Conclusion

The chapter introduced a general framework for the representation of objects by the probability density function over a set of measurements, given by vectors of local characteristics. All degrees of freedom of the object recognition, as introduced in section 1.2 have been incorporated into the model. The estimation and representation of the probability density by multidimensional receptive field histograms have been motivated and discussed.

Section 4.2 developed the analogy between information theory and object recognition. This analogy allows us to use the transinformation for the comparison of different measurement sets. Results indicate that we can predict the quality of recognition by using the transinformation.

The developed statistical representation scheme may be seen as a general framework for the representation of objects. Using multidimensional receptive field histograms shows that the model is a direct generalization of color histograms. The next two chapters therefore use histogram matching techniques for the recognition of objects. Chapter 7 will extend the application of the proposed framework to a probabilistic object recognition algorithm which is capable of recognizing objects in the presence of partial occlusion and in cluttered scenes. Chapter 8 defines two active object recognition algorithms and chapter 9 object classification based on the proposed statistical object representation model.

Chapter 5

Measurements for histogram matching

The previous chapter introduced multidimensional receptive field histograms for the statistical representation of objects. Histogram matching provides a direct application of multidimensional histograms for object recognition. This chapter proposes and analyzes several histogram comparison measurements. The main motivation of histogram matching for object recognition is its low computational cost. Additionally, statistical information for the entire image is used, resulting in robust recognition in the presence changes in object appearance. The next chapter applies the proposed comparison measurements in the context of object identification by histogram matching.

Section 5.1 describes and discusses different comparison measurements in the context of multidimensional receptive field histograms. Section 5.2 examines the stability of the introduced measurements in the presence of additive Gaussian noise, blurring and image plane rotation. Section 5.3 is devoted to the stability analysis of histogram matching with respect to illumination intensity changes. The last section 5.4 concludes with the results of the chapter.

5.1 Histogram comparison measurements

The section describes possible measurements for the comparison of histograms. The analysis of such measurements is important since the *intersection*–measurement, used by Swain and Ballard [Swa 91] in the context of color histograms, has limitations in the more general case of multidimensional receptive field histograms. In particular, the intersection measurement supposes that all measurements are equally probable, which cannot be assumed for an arbitrary measurement set $M = \cup_k m_k$.

Section 5.1.1 introduces the intersection measurement and two possible extensions. Section 5.1.2 describes quadratic distances with the special cases of SSD, sum of squared distances, and the Mahalanobis distance $maha$. Section 5.1.3 motivates the use of χ^2–statistics for the comparison of two histograms. The indexing in large databases (as for example in image retrieval) requires efficient search strategies. Section 5.1.4 briefly discusses how to organize a histogram databases in order to obtain an efficient indexing strategy.

As motivated in section 4.1, an object o_n is represented by the probability density function over a measurement set $M = \cup_k m_k$. Each measurement m_k corresponds to an L–dimensional vector of local descriptors. We use several histograms $H(M|o_n, r_x, r_y, r_z, t_z)$ as the discrete approximation of the probability density function of an object o_n. Each histogram corresponds to a particular appearance of an object, which is defined by a fixed rotation (r_x, r_y, r_z) and scale t_z. In the context of object recognition by histogram matching we compare histograms of the database with the histogram of the actual test image. In the following $V = \cup_{\mathbf{i}} v_{\mathbf{i}}$ refers to a histogram of the histogram database, corresponding to a particular appearance of a database object. The histogram of a test image is signified by $Q = \cup_{\mathbf{i}} q_{\mathbf{i}}$. $\mathbf{i}$ 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).

5.1.1 Intersection measurements

Swain and Ballard introduced in their original work [Swa 91] the *intersection*-measurement for the comparison of color histograms. The intersection of two histograms V and Q is defined by:

$$\cap(Q, V) = \sum_{\mathbf{i}} \min(q_{\mathbf{i}}, v_{\mathbf{i}}) \tag{5.81}$$

The intuitive motivation for this measurement is the calculation of the common part (the intersection) of two histograms V and Q. An advantage of this measurement is that it explicitly neglects background pixels, which may occur in the test histogram Q but do not occur in the database histogram V. This measurement is computational inexpensive since only two operations (minimum and addition) are necessary for each histogram cell. The complexity is linear with the number C of histogram cells: $O(C)$.

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 [Cal 93]. A second inconvenience of the intersection is that all histogram cells are treated equally and should therefore be equally probable. An appropriate scaling of the different axes of the multidimensional histograms can be used (for example logarithmic scaling) but for the sake of stability. Therefore, histogram cells should not be assumed equally probable in the general case of multidimensional receptive field histograms.

In order to overcome the problem that each histogram cell is equally treated, Ennesser and Medioni [Enn 93] improved the intersection measurement by weighting each cell with a factor

$w_{\mathbf{i}}$. The weighting factor $w_{\mathbf{i}}$ should be chosen so as to be proportional to the "importance" (or significance) of a particular color (or histogram cell $\mathbf{i}$). If the average histogram $A = \cup_{\mathbf{i}} a_{\mathbf{i}}$ is known we may use the following weighted intersection measurements:

$$\cap_{we}(Q,V) = \sum_{\mathbf{i}} \frac{1}{a_{\mathbf{i}}} \min(q_{\mathbf{i}}, v_{\mathbf{i}}) \tag{5.82}$$

$$\cap_{mo}(Q,V) = \sum_{\mathbf{i}} \frac{v_{\mathbf{i}}}{a_{\mathbf{i}}} \min(q_{\mathbf{i}}, v_{\mathbf{i}}) \tag{5.83}$$

For $\cap_{we}(Q,V)$ the weighting factor $w_{\mathbf{i}} = \frac{1}{a_{\mathbf{i}}}$ is inversely proportional to the overall occurrence of the histogram cell $\mathbf{i}$. This weighting factor can be seen as an indication of the significance of histogram cell $\mathbf{i}$. In the second case of $\cap_{mo}(Q,V)$, the weighting factor $w_{\mathbf{i}} = \frac{m_{\mathbf{i}}}{a_{\mathbf{i}}}$ takes into account the significance of the histogram cell as well as the number of occurrences in the model histogram V. The application of weighted intersection measurements may improve the ability to discriminate different objects. But the weighting factors make the intersection measurement less stable for histogram cells $\mathbf{i}$ with low values of $a_{\mathbf{i}}$ (the stability of different measurements is examined in section 5.2).

Stricker [Str 92] proposes another measurement $m(Q,V)$ which is equivalent to the intersection measurement, normalized by the size of the histograms. Stricker introduces the following measurement $m(Q,V)$:

$$m(Q,V) = 1 - \frac{d(Q,V)}{s(Q)+s(V)} \tag{5.84}$$

with

$$d(Q,V) = \sum_{\mathbf{i}} ||q_{\mathbf{i}} - v_{\mathbf{i}}|| \tag{5.85}$$

$$s(Q) = d(Q,0) = \sum_{\mathbf{i}} q_{\mathbf{i}} \tag{5.86}$$

It can be shown that this measurement $m(Q,V)$ is equivalent to the intersection measurement, normalized by the sum of the histogram sizes $s(Q)+s(V)$:

$$\begin{aligned}
s(Q)+s(V) &= \sum_{\mathbf{i}} q_{\mathbf{i}} + \sum_{\mathbf{i}} v_{\mathbf{i}} \\
&= 2\sum_{\mathbf{i}} \min(q_{\mathbf{i}}, v_{\mathbf{i}}) + \sum_{\mathbf{i}} ||q_{\mathbf{i}} - v_{\mathbf{i}}|| \\
&= 2\cap(Q,V) + d(Q,V) \\
\Longleftrightarrow s(Q)+s(V)-d(Q,V) &= 2\cap(Q,V) \\
\Longleftrightarrow m(Q,V) &= \frac{2\cap(Q,V)}{s(Q)+s(V)} \qquad \Box
\end{aligned}$$

5.1.2 Quadratic distances

Quadratic distances between two histograms V and Q can be written in matrix notation by using a square matrix W of weights. The matrix W takes into into account relationships between different histogram cells:

$$quad(Q,V) = (Q-V)^T W (Q-V) \tag{5.87}$$

A special case is the sum of squared distances SSD commonly used in signal processing (with W the identity matrix):

$$SSD(Q,V) = \sum_{\mathbf{i}} (q_{\mathbf{i}} - v_{\mathbf{i}})^2 \tag{5.88}$$

A sensible choice of matrix W is the inverse of the covariance matrix of histogram cells, modeling the significance of each cell and the dependencies between different cells. This special case is known as the Mahalanobis distance. When different cells of the histogram are mutually independent only the diagonal elements of the matrix W are nonzero and we obtain a special Mahalanobis distance, which we call $maha(Q,V)$ throughout this thesis:

$$maha(Q,V) = \sum_{\mathbf{i}} \frac{(q_{\mathbf{i}} - v_{\mathbf{i}})^2}{\eta_{\mathbf{i}}^2} \tag{5.89}$$

with $\eta_{\mathbf{i}}^2$ the variance of the histogram cell $\mathbf{i}$.

As mentioned above, equation 5.87 allows us to introduce relationships between different histogram cells. As mentioned above we can use the inverse of the complete covariance matrix as matrix W. Other possibilities for the choice of matrix $W = (w_{\mathbf{i}j})$ include [Haf 95]:

$$w_{\mathbf{i}j} = 1 - \frac{d_{\mathbf{i}j}}{\max_{\mathbf{i}j}(d_{\mathbf{i}j})} \tag{5.90}$$

$$w_{\mathbf{i}j} = e^{-\sigma\left(\frac{d_{\mathbf{i}j}}{\max_{\mathbf{i}j}(d_{\mathbf{i}j})}\right)^2} \tag{5.91}$$

with $d_{\mathbf{i}j}$ the Euclidean distance (or any other distance) between two histogram cells $\mathbf{i}$ and $\mathbf{j}$. By taking into account such simple relationships we can obtain stable results, the price for this stability is a decrease in ability to discriminate between different objects. Whenever the covariance matrix can be estimated reliably, the Mahalanobis distance is preferred. Another interesting idea is the use of the Fisher information matrix as matrix W.

The complexity of quadratic distances $quad$ (equation 5.87) is generally $O(C^2)$ with C the number of histogram cells. The exact complexity depends upon the matrix W. In particular, the overall complexity of SSD (equation 5.88) and the Mahalanobis distance $maha$ (equation 5.89) is linear with the number C of histogram cells: $O(C)$. Nevertheless the computational cost for both is higher than for the intersection measurement, since SSD needs three operations per histogram cell (subtraction, multiplication and addition) and the Mahalanobis distance $maha$ uses four operations per cell (subtraction, multiplication, division and addition).

5.1.3 χ^2 – statistics

The formal statistical method for determining if two distributions differ is the χ^2–test. Starting from the null hypothesis that two data sets (histograms) are drawn from the same population (for example measurements of the same object) the goal is to disprove the hypothesis. Disproving the hypothesis proves that the histograms are drawn from different distributions. Failing to disprove the null hypothesis only shows that the two histograms are consistent and could be drawn from the same population. Therefore, the χ^2–test is a consistency test for two histograms.

In the context of multidimensional receptive field histograms we use χ^2–statistics without using the χ^2–probability function. χ^2–statistics are used for the calculation of "dissimilarity" between two histograms. Two different ways of calculation of the χ^2–statistics are considered in this work [Pre 92]. The first – $\chi^2_v(Q,V)$ – assumes exact knowledge of the model histogram V:

$$\chi^2_v(Q,V) = \sum_{\mathbf{i}} \frac{(q_{\mathbf{i}} - v_{\mathbf{i}})^2}{v_{\mathbf{i}}} \tag{5.92}$$

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

$$\chi^2_{qv}(Q,V) = \sum_{\mathbf{i}} \frac{(q_{\mathbf{i}} - v_{\mathbf{i}})^2}{q_{\mathbf{i}} + v_{\mathbf{i}}} \tag{5.93}$$

As we will later conclude from experiments, these two χ^2–statistics provide better recognition results for most cases than all other measurements. Unfortunately, neither of the χ^2–statistics is a metric since the triangulation inequality is not satisfied. In order to see this we can consider the degenerate case of one-cell histograms $A = (a)$ and $C = (c)$. For any one-cell histogram $B = (b)$ with $a < b < c$ the following inequality holds (which is the opposite of the usual triangular inequality of a metric):

$$\chi^2_{qv}(A,B) + \chi^2_{qv}(B,C) < \chi^2_{qv}(A,C) \tag{5.94}$$

Since the χ^2–statistics give the best recognition results we introduce a modified measurement, which is a metric. Supposing knowledge of the average histogram $A = \cup_{\mathbf{i}} a_{\mathbf{i}}$ we can use the following equation:

$$\chi^2_{av}(Q,V) = \sum_{\mathbf{i}} \frac{(q_{\mathbf{i}} - v_{\mathbf{i}})^2}{a_{\mathbf{i}}} \tag{5.95}$$

Unfortunately this modified measurement does not give the same quality of result as the two original χ^2–statistics, even though it provides good results (relative to other comparison measurements).

The overall complexity of all three χ^2–statistics is $O(C)$ with C the number of histogram cells. The number of operations per histogram cell is five in the case of χ^2_{qv} and four in the case of χ^2_v and χ^2_{av}.

5.1.4 Efficient histogram indexing

For the measurements that we have introduced, the computational complexity for the comparison of two histograms is typically $O(C)$ with C the number of histogram cells (a complexity of $O(C^2)$ is obtained for the quadratic distance *quad*, which is not considered in the following discussion). Table 5.1 shows typical frequencies obtained by 5 different comparison measurements for the matching of histograms with 4096 cells. The table shows that between 500 and 2000 histograms can be compared per second.

measurement	$\cap$	χ^2_v	χ^2_{qv}	SSD	$maha$
frequency [Hz]	1827	1241	845	1767	623

Table5.1: Typical frequencies for the comparison of histograms with 4096 cells. The frequencies have been obtained for the experiment of section 6.3 on a Silicon Indy 200 MHz

Besides the complexity (and therefore the time) for the comparison of two histograms, it is also important to discuss the indexing complexity for large databases. This is an important issue for example in the context of current research into image retrieval by image contents. Image retrieval is motivated by the fact that it is impractical (and time consuming) to store manually added keywords for each image of a database. A prominent candidate for image retrieval is the color indexing scheme of Swain and Ballard, which can be found (in its original or in an improved version) in many image retrieval systems.

By calculating for each image within a database, a multidimensional histogram, image retrieval becomes an indexing problem into a histogram database. Supposing a complexity of $O(C)$ for the comparison of two histograms with C cells and applying a linear search on a database of N histograms the "indexing" complexity is $O(C \times N)$. This complexity is unacceptable for large databases. It is usually desirable to obtain a logarithmic database search of complexity $O(C \log(N))$. Principally, two types of approach exist:

- Organization of the histogram database in a tree structure (or similar structure).
- Indexing of the histogram database with a hash table. The main issue here is finding a suitable index.

The first approach, the organization of the histogram database in a tree–like structure, typically relies on the fact that the comparison measurement is a metric (which is not the case for the χ^2–statistics, see section 5.1.3). Assuming a metric comparison measurement, many algorithms exist for the organization of databases in tree–like structures.

The second possibility is more interesting, as an index for a histogram should capture important information without being unstable. An example is the application of moments of histograms as indices, as is the case with the average of a histogram [Haf 95]. Any dimensionality reduction algorithm may be applied for the extraction of indices, for example linear discriminant analysis or the Karhunen–Loeve transformation. Once again the underlying assumption is that the comparison measurement is a metric.

Using a quadratic distance or an intersection measurement, both approaches can be applied. Many standard algorithms exist which are not discussed here in more detail. Unfortunately the χ^2–statistics are not metric (see section 5.1.3), which makes the application of such standard algorithms difficult. Therefore, we propose a specific algorithm for the organization of histogram databases. The idea is to calculate a histogram pyramid for each histogram of the database. More precisely, the first histogram of the pyramid is calculated directly on the image with the desired maximal histogram resolution. The histogram of the next level is then calculated on the basis of the first histogram by reducing the resolution for each histogram axis by a given factor (for example by a factor of 2). The histogram pyramids can be calculated off-line. During matching a branch and bound algorithm may be applied to the histogram pyramids of the database. For the application of such an algorithm lower and upper bounds are necessary. The computational cost of such an algorithm is dependent strongly on the quality of these bounds. Since the quality of such bounds cannot be assured for a particular database it is meaningless to calculate the average complexity of the resulting algorithm. Nevertheless we will give a lower and an upper bound, which are exact in the sense that the histogram with minimal χ^2_{qv} can be found. The bounds are also optimal, since no smaller upper bound and no larger lower bound exist.

Lower bound for χ^2_{qv}:

In order to apply a branch and bound algorithm on a histogram pyramid a lower bound is necessary. Without losing generality, we consider two histograms with only two cells each: $A = a_0 \cup a_1$ and $B = b_0 \cup b_1$. For each of the histograms we calculate the next level of its "pyramid" which is given by: $A' = [a_0 + a_1]$ and $B' = [b_0 + b_1]$. Both histograms A' and B' contain one single cell. The essential question is how to obtain a lower bound for the comparison of A and B by the comparison of A' and B'. We will show that the lower bound is given by:

$$\chi^2(A', B') \leq \chi^2(A, B) \tag{5.96}$$

We need to establish that:

$$\chi^2([a_0 + a_1], [b_0 + b_1]) \leq \chi^2(a_0, b_0) + \chi^2(a_1, b_1)$$

Proof:

$$\begin{aligned}
\frac{(a_0 + a_1 - b_0 - b_1)^2}{a_0 + a_1 + b_0 + b_1} &\leq \frac{(a_0 - b_0)^2}{a_0 + b_0} + \frac{(a_1 - b_1)^2}{a_1 + b_1} \\
\frac{((a_0 - b_0) + (a_1 - b_1))^2}{(a_0 + b_0) + (a_1 + b_1)} &\leq \frac{(a_0 - b_0)^2(a_1 + b_1) + (a_1 - b_1)^2(a_0 + b_0)}{(a_0 + b_0)(a_1 + b_1)}
\end{aligned}$$

With the following substitutions: $\alpha = a_0 - b_0, \beta = a_1 - b_1, \gamma = a_0 + b_0$ and $\delta = a_1 + b_1$ we obtain:

$$\begin{aligned}
(\alpha + \beta)^2\gamma\delta &\leq (\alpha^2\delta + \beta^2\gamma)(\gamma + \delta) \\
2\alpha\beta\gamma\delta &\leq \beta^2\gamma^2 + \alpha^2\delta^2 \\
2(a_0^2 - b_0^2)(a_1^2 - b_1^2) &\leq (a_1 - b_1)^2(a_0 + b_0)^2 + (a_0 - b_0)^2(a_1 + b_1)
\end{aligned}$$

For $a_i + b_i = 0$ for $i = 1$ or $i = 2$, the left hand side becomes zero. For $a_i + b_i > 0$ for $i = 1, 2$ we obtain:

$$\begin{aligned} 8a_0a_1b_0b_1 &\leq 4a_1^2b_0^2 + 4a_0^2b_1^2 \\ 0 &\leq 4(a_1b_0 - a_0b_1)^2 \qquad \square \end{aligned}$$

Remark: in the case of $a_0 = b_0 = 0$, the equality of equation 5.96 is satisfied, implying that no larger lower bound exists.

Upper bound for χ^2_{qv}:

The upper bound is given by (using $s(A) = \sum a_i$ as size of histogram A):

$$s(A') + s(B') \geq \chi^2_{qv}(A, B) \tag{5.97}$$

We need to establish that

$$a_0 + a_1 + b_0 + b_1 \geq \chi^2(a_0, b_0) + \chi^2(a_1, b_1)$$

which is given directly by ($i = 1, 2$):

$$\begin{aligned} (a_i + b_i)^2 &\geq (a_i - b_i)^2 \\ (a_i + b_i) &\geq \frac{(a_i - b_i)^2}{a_i + b_i} \qquad \square \end{aligned}$$

Remark: in general no smaller upper bound exists since the equality of equation 5.97 can be obtained in the case of $a_1 = b_0 = 0$.

5.2 Stability of histogram comparison measurements

The section examines the stability of different comparison measurements introduced in the previous section 5.1. The stability of the measurements with respect to changes in the histograms is a necessary condition for the usefulness of histogram matching for recognition. Such changes in a histogram may be caused either by changes in object appearance or by different sources of perturbation such as image noise and image blurring.

In our experiments we compare eight different measurements:

- three χ^2–statistics (section 5.1.3): χ^2_v, χ^2_{qv} and the modified χ^2–statistics χ^2_{av},
- two quadratic distances (section 5.1.2): SSD and $maha$ and
- three intersection measurements (section 5.1.1): $\cap$, $\cap_{we}$ and $\cap_{mo}$.

In order to make the different measurements comparable we calculate a normalized measurement d_{norm}. The normalization uses the optimal value *opt* and the average value *av* of each measurement. The average value is calculated on the basis of a large number of histograms (378 for the image plane rotation and 500 for the noise and the blurring experiments). Knowing *opt* and *av* we normalize the calculated measurement d_{raw} between two histograms with the following equation:

$$d_{norm} = \frac{d_{raw} - opt}{av - opt} \times 0.5 \tag{5.98}$$

Therefore, $d_{norm} = 0.0$ corresponds to the optimal value *opt* of a measurement and $d_{norm} = 0.5$ corresponds to the average value *av*.

We have seen in recognition experiments that a value of $d_{norm} \leq 0.1$ is usually sufficient for the identification of objects. In the presence of significant histogram changes it is difficult to achieve such a low value. A value of $d_{norm} \leq 0.2$ has been shown to be an appropriate value in order to obtain the correct match within a small number of matches (for example within the first 5% of the database). In the following stability analysis, we use these two values ($d_{norm} \leq 0.1$ and $d_{norm} \leq 0.2$) as reference values. We should mention that their exact values do not influence the analysis significantly, since we will examine the relative stability between different measurements rather than "absolute" stability. More important than the absolute stability is the ability to discriminate different objects by histogram matching. Chapter 6 is devoted to this topic.

We can interpret pixel values (grayvalues or colorvalues) as the most local (and least stable) characteristics of an image. Since we want to analyze the stability of different histogram comparison measurements we have chosen the histogram of the local characteristic *grayvalue* as a suitable – meaning unstable – basis. This choice is also motivated by the wish to separate the stability introduced by comparison measurements from the stability which due to robust local characteristics as for example Gaussian derivatives and Gabor filters. In the following we show results for two different resolutions of grayvalue histograms, namely for full resolution of 256 cells per histogram (each histogram cell corresponds to one grayvalue, always shown left in the figures) and with reduced resolution of 8 cells per histogram (each histogram cell corresponds to 32 grayvalues, always shown right in the figures).

Throughout the section we present the results as graphs. For each of the examined changes (additive Gaussian noise, blurring and image plane rotation) we show six graphs in figures 5.2, 5.4 and 5.6. The left column in these figures corresponds to grayvalue histograms of full resolution (resolution of 256 cells) and the right column corresponds to grayvalue histograms of reduced resolution: 8 cells per histogram. The two graphs shown in the same row of these figures correspond to a particular set of comparison measurements. The vertical axis of these graphs show the normalized measurement d_{norm} for the interval $[0.0, 0.5]$. As a reference we always include the graph of "changes in the histogram", corresponding to the effective histogram changes (for example in figure 5.2 (a) the reference graph shows an absolute change of the histogram between 0.0 (corresponding to 0% change) and 0.35 (corresponding to a change of 35% of the histogram).

The following section 5.2.1 examines stability with respect to additive Gaussian noise and section 5.2.2 in relation to blurring. Section 5.2.3 examines the stability in the presence of image plane rotation. We use the following notation in the figures of this section:

- χ^2–statistics: chsone for χ^2_v, chstwo for χ^2_{qv} and chsav for χ^2_{av} (see section 5.1.3),
- intersection measurements: intersection for $\cap$, inter_We for $\cap_{we}$ and inter_Mo for $\cap_{mo}$ (see section 5.1.1),
- quadratic distances: SSD for SSD and maha for $maha$ (see section 5.1.2).

5.2.1 Stability with respect to additive Gaussian noise

In order to examine the stability of different comparison measurements with respect to additive Gaussian noise we use 500 images of objects randomly chosen from our image database. To each pixel value of these images we add Gaussian noise with zero mean and standard deviation $\sigma = 0.5, 1, 1.5, \ldots 20$. Figure 5.1 shows an image with added Gaussian noise for $\sigma = 5, 10, 15$ and 20. In the experiment the histogram of the original image (without additive Gaussian noise) is compared with histograms of "noisy" images. In the graphs of figure 5.2 we plot the standard deviation σ of the additive Gaussian noise (horizontal axis) against the normalized measurement d_{norm} (see equation 5.98). Each point in these graphs corresponds to the average normalized measurement obtained over 500 noisy images. As mentioned above, the graph of the absolute changes in the histograms relative to the original histogram (averaged over 500 images) is plotted as reference.

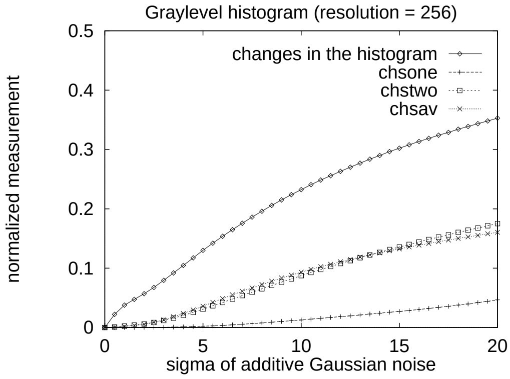
Figure5.1: Stability with respect to additive Gaussian noise: one of the 500 images without additive Gaussian noise (the leftmost image) and with Gaussian noise of $\sigma = 5, 10, 15, 20$ (from left to right)

A first result is that the histograms of resolution 8 give more stable results than the histograms with resolution 256 independent of the comparison measurement (see figure 5.2). This result seems to be obvious but the result is obtained not only for the absolute measurements but for the normalized measurements d_{norm}. In other words, in the case of resolution 8 the values of the comparison measurements change less relative to the average measurement av, which is calculated separately for resolution 8 and 256. This result indicates that histograms with lower resolution are more stable to additive Gaussian noise.

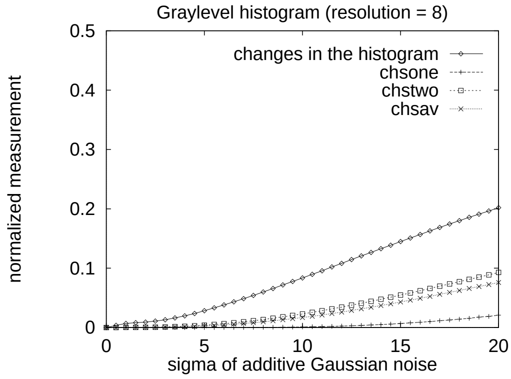
The first row of figure 5.2 shows results of the χ^2–statistics. Stable results are obtained for all three χ^2–statistics since the normalized measurement d_{norm} stays below 0.2 all the time and below 0.1 even for large σ. In particular, χ^2_v gives the best results compared with all eight measurements.

The second row of figure 5.2 shows graphs of the three intersection measurements. The results are stable (for the resolution of 256 the measurements stay below 0.1 for $\sigma \leq 5$), even though significantly worse than the results of the χ^2–statistics. We can rank $\cap_{we}$ superior to the other two intersection measurements $\cap$ and $\cap_{mo}$.

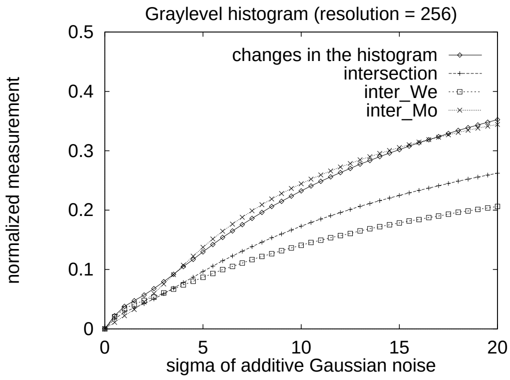
The third row of figure 5.2 shows results of the two quadratic distances, namely the sum of squared differences SSD (equation 5.88) and the Mahalanobis distance $maha$ (equation 5.89). SSD shows a high stability to additive Gaussian noise (comparable with χ^2_{qv} and χ^2_{av}). The Mahalanobis distance also shows a stable behavior for small values of σ. For large values of σ the Mahalanobis distance becomes unstable particularly for the high resolution case of 256 (figure 5.2 (e)). A high stability is obtained in the low resolution of 8. The observed degradation



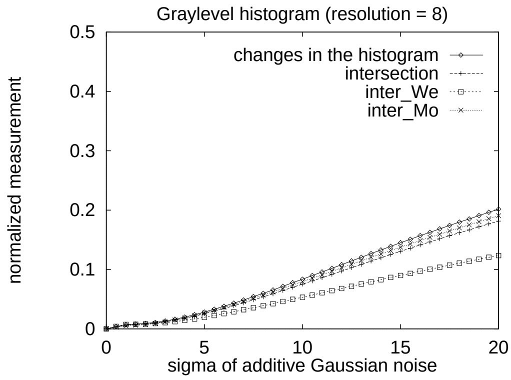
(a)



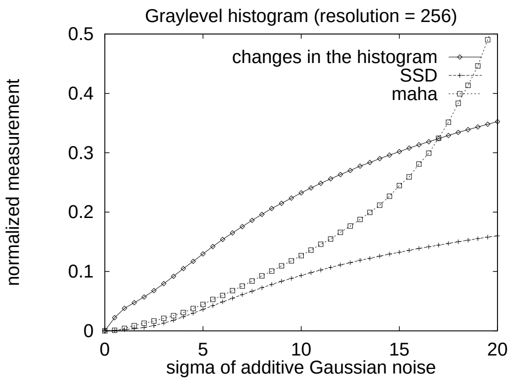
(b)



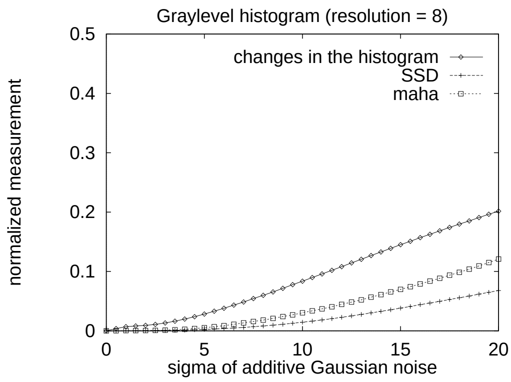
(c)



(d)



(e)



(f)

Figure5.2: Stability with respect to additive Gaussian noise. For comments see section 5.2.1

for the resolution of 256 is due to the fact that we did not consider Gaussian noise during the estimation of the variances $\eta_{\mathbf{i}}^2$ (which are needed for the calculation of the Mahalanobis distance, see equation 5.89). The estimation of variances is always critical for the application of the Mahalanobis distance, since the calculation of the Mahalanobis distance includes the division by the estimated variances. Therefore, the Mahalanobis distance is typically less stable than the sum of squared distances SSD.

From this first experiment we can conclude that χ_v^2 provides the most stable results followed by χ_{qv}^2, χ_{av}^2 and SSD. The Mahalanobis distance may be stable, if the variances can be estimated robustly. The three intersection measurements provide acceptable results, even though less stable than the other measures.

5.2.2 Stability with respect to blurring

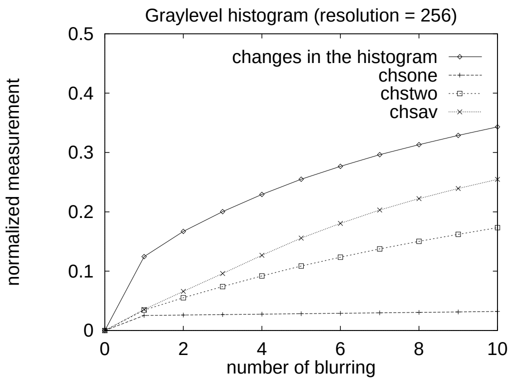
A second experiment examines the stability of histogram comparison measurements in relation to repeated mean filtering with a 3×3 mask. Mean filtering is used as a blurring operation, which can be seen as a simulation of image blur due to defocusing the camera. In order to increase the blurring effect we repeat the mean filtering several times for the same image. Figure 5.3 shows an example image and different blurred version of the image. As in the previous experiment we use 500 images of objects of our image database.

Figure5.3: Stability with respect to blurring: one of the 500 images where a 3×3 mean-filter has been applied $2, 4, 7$ and 10 times

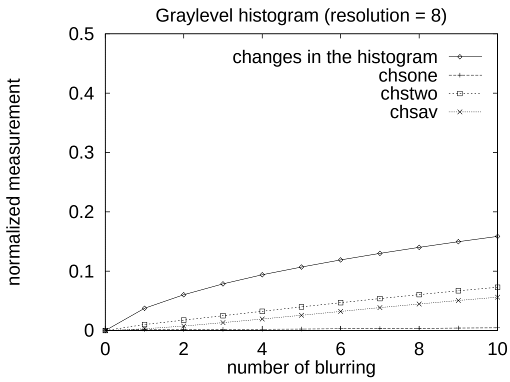
As introduced above we use the normalized measurement d_{norm} in order to make the different measurements comparable. In the graphs of figure 5.4, the vertical axis corresponds to the normalized measurement d_{norm} and the horizontal axis shows the number of repeated mean filtering (blurring operation).

The graphs 5.4(a) and (b) show results of the χ^2–statistics. As in the first experiment, χ_v^2 gives the most stable results relative to the other measurements. The results of χ_{qv}^2 are stable for the resolution of 8 and reasonably stable (overall the second best results) for resolution of 256. This time χ_{av}^2 provides less stable results than χ_{qv}^2 for the resolution of 256. The results of χ_{av}^2 for the resolution of 8 are slightly better than the results of χ_{qv}^2.

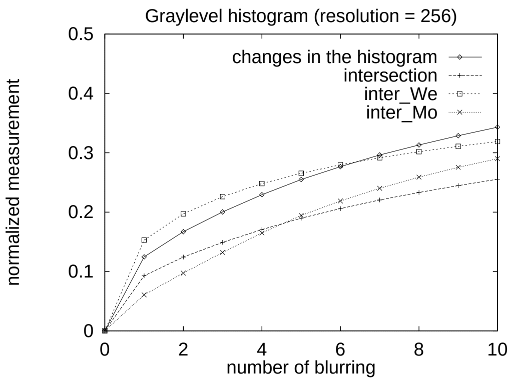
The graphs 5.4(c) and (d) show results of the intersection measurements. In the case of resolution 256 a single mean filtering causes a significant increase in the value of the different intersection measurements. This is due to the change of the histogram which is shown by the graph "changes in the histogram". As a matter of fact, the intersection measurements depend directly on these changes. In both graphs 5.4(c) (resolution = 256) and (d) (resolution = 8) the two measurements $\cap$ and $\cap_{mo}$ provide very similar results, whereas $\cap_{we}$ gives worse results.



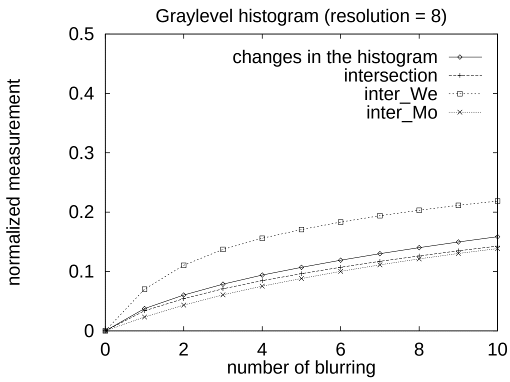
(a)



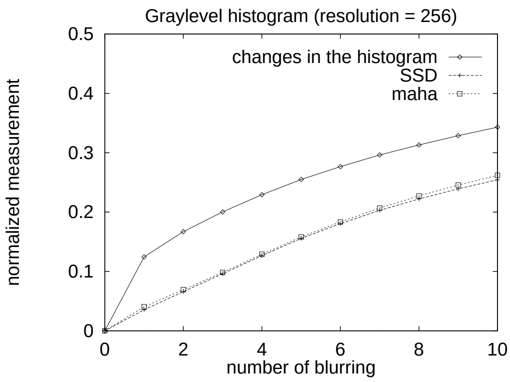
(b)



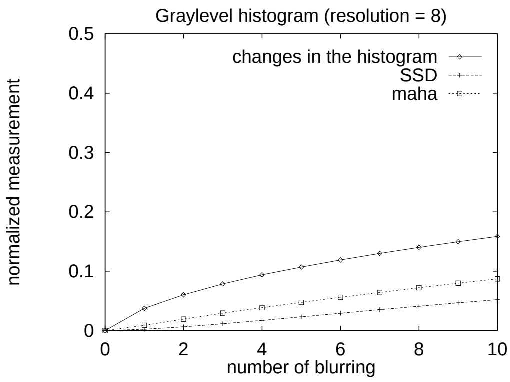
(c)



(d)



(e)



(f)

Figure5.4: Stability with respect to blurring. For comments see section 5.2.2

The graphs 5.4(e) and (f) show results of the quadratic distances. Here the two distances behave similarly. The obtained stability is high for the low resolution of 8 and comparable to the results of χ^2_{qv} and χ^2_{av}. For the high resolution of 256 the normalized measurement d_{norm} is below 0.1 until 3 times mean filtering and below 0.2 until 7 times mean filtering. These results are better than for the intersection measurements but not as good as the results obtained by χ^2_v.

In summary we can state that the χ^2–statistics provide the most stable results. More precisely χ^2_v gives the best results. The next best results are obtained by χ^2_{qv}, χ^2_{av} and SSD. $maha$ gives slightly worse results. The worst results are those given by the intersection measurements, where $\cap$ and $\cap_{mo}$ usually obtain better results than $\cap_{we}$.

5.2.3 Stability with respect to image plane rotation

This section evaluates the stability of grayvalue histograms to image plane rotation. Theoretically, grayvalue histograms are invariant with respect to image plane rotation (using a circular support for the calculation of the histograms). But due to noise and the rectangular pixel shape, the observed changes of these histograms are significant. The graph of the "changes of the histogram" in figure 5.6 (a) show that the changes (for resolution = 256) are between 13% and 20%. For the resolution of 8 (figure 5.6 (b)) these changes are less significant (below 10%). In the case of high resolution the changes correspond to an additive Gaussian noise between approximately $\sigma = 5$ and 8.

Figure5.5: 5 rotations for one of 21 objects

In this experiment we use 18 rotated images of 21 objects. Figure 5.5 shows 5 rotations of one object. The rotations are approximately 20° apart. The graphs of figure 5.6 show the approximate rotation angle between a reference image and a test image of the same object. Each of the 18 images of an object have been taken as reference images. Each point in the graph therefore corresponds to an average over $18 \times 21 = 378$ images.

The first row of figure 5.6 shows the results obtained for the χ^2–statistics. In the case of high resolution the results of χ^2_v are better than the results of χ^2_{qv}. Here χ^2_{av} gives less stable results (even though the next best results together with SSD and $maha$). In the case of the low resolution of 8, χ^2_v overall gives the best results whereas the other two measurements give slightly worse results.

The second row of figure 5.6 shows that the intersection measurements give unstable results. The best intersection measurement is the unweighted intersection measurement $\cap$. $\cap_{mo}$ gives worse results and $\cap_{we}$ gives the less stable results of all measures.

The third row of figure 5.6 shows that the results of SSD and $maha$ are comparable to the results of χ^2_{av} for the high resolution of 256. In the case of the low resolution of 8 only SSD gives good results (comparable with χ^2_{av} and χ^2_{qv}). In this low resolution case the Mahalanobis

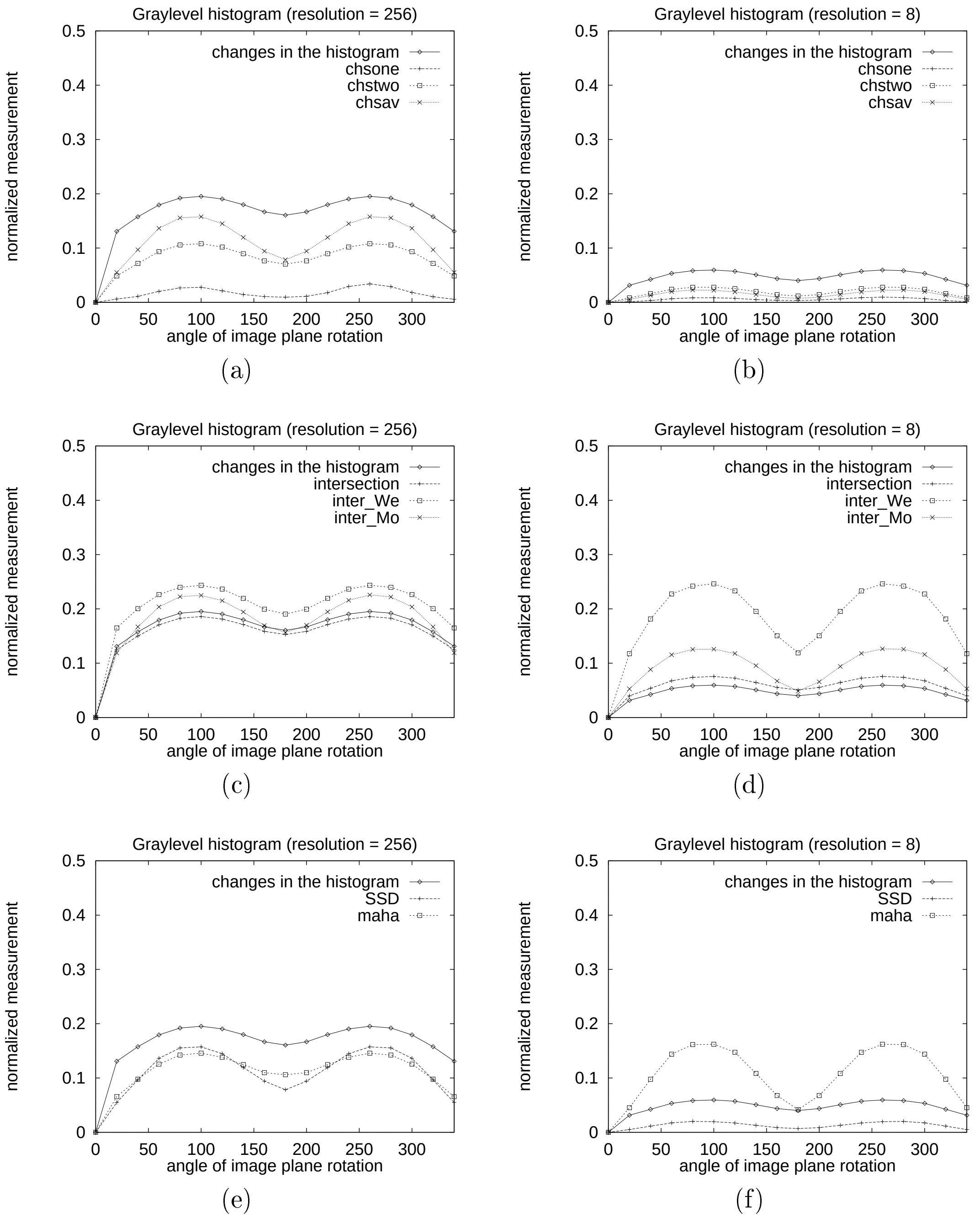


Figure5.6: Stability with respect to image plane rotation. For comments see section 5.2.3

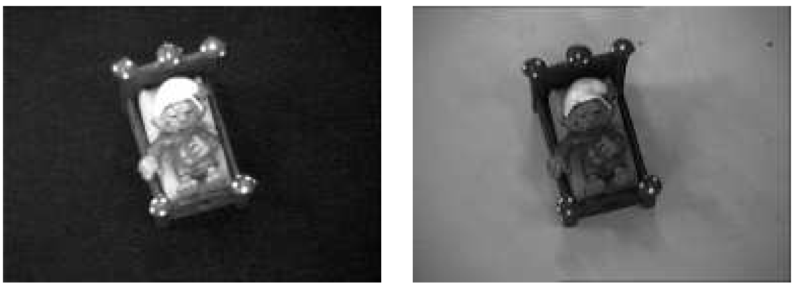
distance *maha* gives the worst results even though the normalized measurement d_{norm} stays below the reference value 0.2.

The conclusion of this experiment is similar to the results of the first two experiments (sections 5.2.1 and 5.2.2): χ_v^2 gives the most stable results, followed by SSD, χ_{av}^2 and χ_{qv}^2. The Mahalanobis distance is not as good as the first four measurements. The intersection measurements provide the worst results ($\cap$ better than $\cap_{mo}$, which is better than $\cap_{we}$).

5.3 Stability of histogram matching with respect to changes in ambient illumination

In what follows we describe experiments that examine the stability of histogram matching with respect to intensity changes in ambient illumination. As in the previous section 5.2 we want to evaluate different comparison measurements. In contrast to the previous section it is meaningless to use grayvalue histograms which change dramatically in the presence of illumination intensity changes. Therefore, we use histograms of Gaussian derivatives. Additionally, we apply different normalization techniques to the filter output. In particular, we compare *energy*, *variance–mean* and *no* normalization of the filter output (see for details of the normalization techniques section 3.2).

The influence and the success of the normalization techniques depends strongly on the characteristics of the camera used. For the majority of our experiments we used the Canon VCC1 camera. As an analysis [Col 96] showed, the automatic gain control of this camera maintains the average gray-level as well as the grayvalue variance constant. Illumination intensity changes for the entire image are therefore not measurable as changes in the recorded image. As long as illumination intensity changes are constant over the entire image, no normalization of filter output is necessary. Nevertheless, since the gain control of the camera is global on the entire image, local intensity changes can be caused for example by a background change as illustrated in figure 5.7.



(a) (b)

Figure5.7: Two images of the same object in front of two different backgrounds. Important local intensity changes can be observed

In order to obtain a large bandwidth of illumination intensity changes we use here a digital Indy-Cam, delivered with the Silicon Graphics Indy. Typically we avoided using his camera due to its poor quality. Since the automatic gain control can be turned off, the camera has however been appropriate for this experiment. The camera characteristics of the Indy-Cam are shown in figure 5.8. The graph plots the grayvalue variance relative to the average gray-level for three images series.

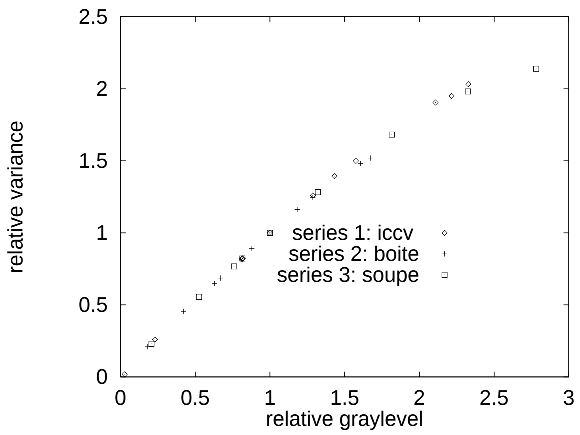


Figure5.8: The camera dynamics of the Indy-Cam (digital Silicon Graphics camera)

The used image series of three objects are shown in figure 5.9 as well as in figures B.1 and B.2 of the appendix. Since the results for all three image series are very similar we describe the results for just a single series (see figure 5.9). During recording of the image series, the illumination intensity of an adjustable halogen lamp was continuously increased. An image with average illumination intensity was chosen as reference image. The numbers below each image of figure 5.9 show the relative gray-level (averaged over the entire image) relative to the reference image.

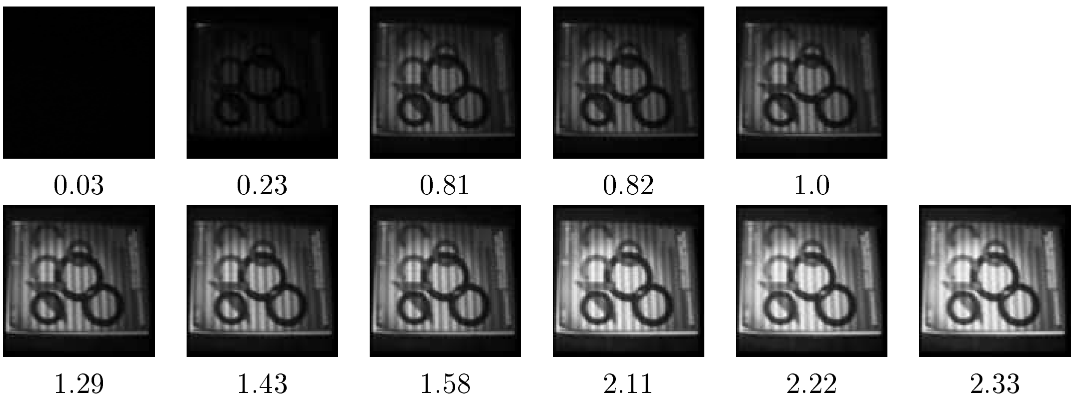


Figure5.9: Image series with illumination intensity changes. The number below each image corresponds to the relative gray-level (averaged over an image)

As mentioned above, the gray-level histograms used in the previous section are not appropriate in the case of illumination intensity changes. This section therefore uses two–dimensional histograms for the filter combination *Dx-Dy* (first Gaussian derivative in x– and in y–direction with $\sigma = 2.0$). The resolution has been 32 cells per histogram axis. Such histograms are typical for the experiments described in chapter 6.

Figure 5.10 shows the application of several histogram comparison measurements. The horizontal axis show the relative gray-level averaged over the entire series of images shown in figure 5.9. The vertical axis shows the normalized measurement as previously introduced (see equation 5.98). In order to estimate the average distance av we used 500 reference images of objects. More precisely figure 5.10 (a) shows the application of χ^2_q (chsone), (b) of χ^2_{qv} (chstwo), (c) of χ^2_{av} (chsav), (d) of $\cap$ (intersection), (e) of SSD (ssd), and (f) of $maha$ (maha) (see section 5.1 for introduction of the measurements).

Each of the graphs of figure 5.10 shows the application of *no* (none), *energy* (eg) and *variance–mean* (vg) – normalization. For no normalization of the filter output the best results are obtained by the χ^2–statistics χ^2_q followed by the results for χ^2_{qv}, χ^2_{av}, SSD and $maha$. The worst results are obtained by the intersection measurement $\cap$. The results of energy normalization and variance–mean normalization are practically identical. For these two normalizations the results of χ^2_q, χ^2_{qv}, χ^2_{av}, SSD and $maha$ are very similar whereas the results of the intersection measurement $\cap$ are again worse.

For all six comparison measurements, energy and variance-mean normalization stabilize the results significantly relative to no normalization of the filter responses. Besides the comparison measurement $\cap$ the maximal distance mostly stays below 0.1, which shows the high stability of energy and variance-mean normalization with respect to illumination intensity changes.

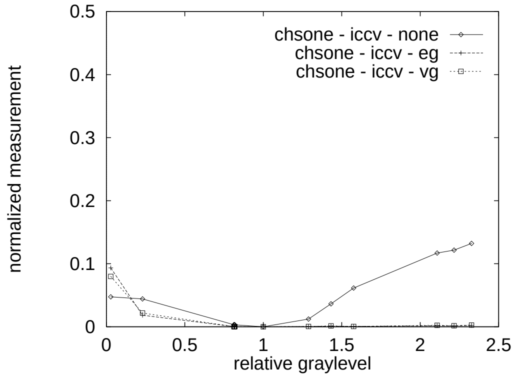
The worst results are obtained in the case of low intensity images (relative intensity of 0.03 and 0.23). This can be explained by the low signal to noise ratio in such images. For all other images the normalized measurement is never far from zero (besides the intersection measurement $\cap$ which is low but slightly worse).

The results described here are typical for this type of experiment. Figures B.1 and B.2 of the appendix show similar results obtained for two other image series. The main conclusion of the experiment is that stable results with respect to illumination intensity changes can be obtained by normalizing the filter output either by energy or by variance–mean. The normalization techniques become unstable in the case of very low intensity images. Probably a similar instability will be obtained in the case of over-saturated images, which we didn't examine here. Besides the intersection measurement all comparison measurements provide stable results and show a similar behavior.

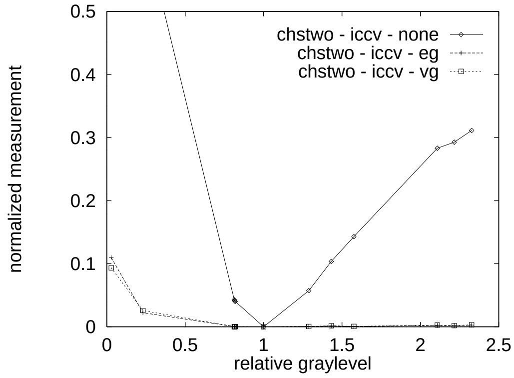
5.4 Conclusion

In this chapter we have introduced different measures for the comparison of histograms. The analysis of these measures is important since the intersection measurement (originally used by Swain and Ballard for the comparison of color histograms [Swa 91]) has theoretical limitations for application in the context of multidimensional receptive field histograms: in order to obtain optimal separation between different histograms the measurements m_k should be equally and sparsely distributed for a particular object. In addition to these theoretical requirements we observe an unstable behavior of the intersection measurement in the presence of additive Gaussian noise, blurring and image plane rotation. Also, in the presence of illumination intensity changes the intersection measurement provided the worst results.

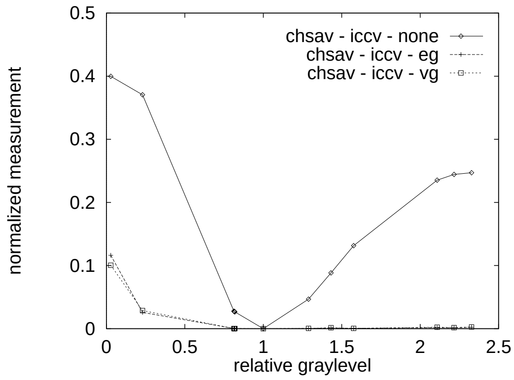
In order to overcome the limitations of the intersection measurements we have proposed several measurements in two classes: χ^2–statistics and quadratic distances. In particular, χ^2–statistics (the method proposed by statisticians for the comparison of histograms) provide stable



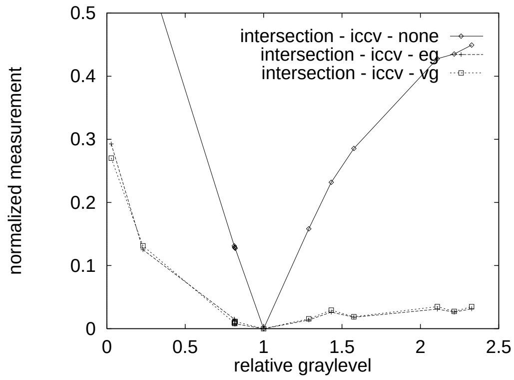
(a)



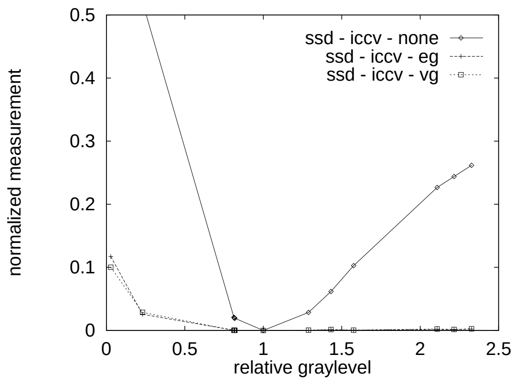
(b)



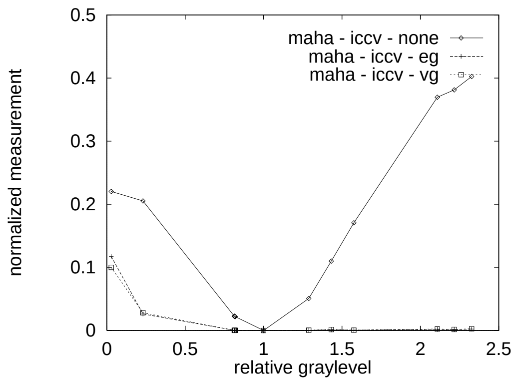
(c)



(d)



(e)



(f)

Figure5.10: Stability with respect to illumination intensity changes for image series shown in figure 5.9: (a) chsone: χ^2_q, (b) chstwo: χ^2_{qv}, (c) chsav: χ^2_{av}, (d) intersection: $\cap$, (e) ssd, (f) maha: Mahalanobis distance

results relative to the examined changes. Quadratic distances allow us to introduce relationships between different histogram cells, which is especially interesting in the presence of significant histogram changes. We used two versions of the quadratic distance, namely the sum of squared distances SSD and the Mahalanobis distance $maha$, which do not take into account relationships between histograms cells. Both measurements showed more stable behavior than the original intersection measurement.

From the experiments of section 5.2 we can conclude (stability analysis with respect to additive Gaussian noise, blurring and image plane rotation):

- χ^2_v obtains the most stable results,
- χ^2_{qv}, χ^2_{av} and SSD provide similar results, which are also stable. In particular circumstances χ^2_{av} becomes less stable,
- the Mahalanobis distance $maha$ gives similar results to SSD but is usually less stable. Particularly in the case of image plane rotation where the results are less stable,
- the three intersection measurements provide the least stable results.

Section 5.3 describes a stability analysis of histogram matching with respect to illumination intensity changes. Results underline the stability of χ^2–statistics and the quadratic distances. The intersection measurement $\cap$ provided less stable results. Section 5.3 examines not only the stability of histogram comparison measurements but also the stability of different normalization techniques. Experiments show that stable results can be obtained in the presence of illumination intensity changes, using either energy or variance–mean normalization.

This chapter analyzed the complexity and the stability of different histogram comparison measurements. More important is the ability of a measurement to discriminate different objects in the context of recognition. The next chapter therefore shows recognition experiments for histogram matching in the presence of changes in the object appearance. Appearance changes include for example scale change, image plane rotation and view point changes.

Chapter 6

Object recognition by histogram matching

Histogram matching is a direct application of multidimensional receptive field histograms for object recognition. Chapter 5 introduced different histogram comparison measurements and analyzed their stability with respect to different changes. This chapter describes experiments for object identification by histogram matching in the presence of image plane rotation, scale changes, viewpoint changes and partial occlusion.

Throughout the chapter we use the following Gaussian derivatives (see section 3.1.1):

- *Dx*: first Gaussian derivative in x–direction
- *Dy*: first Gaussian derivative in y–direction
- *Lap*: Laplacian operator
- *G12*: rotation invariant Gaussian filter based on a mixture of first and second order Gaussian derivatives

Different combinations of these filters are used: two rotation variant filter combinations (*Dx-Dy* and *Dx-Dy-Lap*) and two rotation invariant filter combinations (*Mag-Lap* and *Mag-G12-Lap*). In the following we examine the object recognition performance of different filter combinations as well as the influence of different histogram resolutions. The notation *Dx-Dy-32* for example refers to a histogram of the filter combination *Dx-Dy* with resolution of 32 cells per histogram axis. The corresponding two–dimensional histogram contains $32^2 = 1024$ cells. Due to discretization, each cell corresponds to a particular range of output values of the two dimensional filter combination *Dx-Dy*. Such a range of filter output is called "measurement vector" m_k.

As in the previous chapter, we use the following notations in the figures for the different histogram comparison measurements:

- χ^2–statistics: chsone for χ^2_v, chstwo for χ^2_{qv} and chsav for χ^2_{av} (see section 5.1.3),
- intersection measurements: intersection for $\cap$, inter_We for $\cap_{we}$ and inter_Mo for $\cap_{mo}$ (see section 5.1.1),
- quadratic form distances: SSD for SSD and maha for $maha$ (see section 5.1.2).

The next section 6.1 presents a recognition example for 261 objects by histogram matching. Section 6.2 describes how to account for image plane rotation and section 6.3 considers scale changes. Two experiments of section 6.4 provide recognition results in the presence of image plane rotation and scale changes. The robustness of histogram matching with respect to viewpoint changes is examined in section 6.5. An experiment with a database of 103 objects considering viewpoint changes, image plane rotation and scale changes is described in section 6.6. In particular, this latter section examines the robustness of histogram matching to partial occlusion. A short discussion of the memory requirement for multidimensional receptive field histograms is given in section 6.7. Section 6.8 concludes the results of the chapter.

6.1 An example of object identification by histogram matching

In order to show the applicability of histogram matching for object identification, an experiment on a database of 261 objects is described. The calculation of the histogram database uses one image per object corresponding to a particular appearance of the object. As a test–set we use different images of the same 261 objects, where changes in the appearance of the objects have been introduced. In particular, the scale and the viewpoint have been changed. The database is composed of the following objects, which can be found in appendix A:

- 61 objects from our own image database — see figures A.1 and A.3
- 100 objects from the Columbia image database — see figure A.6
- 100 objects from aerial images of Marseille — see figures A.4 and A.5

The image database contains one single image per object. For each image we calculate a two–dimensional histogram *Dx-Dy-32* : first Gaussian derivatives in x– and y–direction with resolution 32 per histogram axis. Each histogram therefore contains $32^2 = 1024$ cells. These histograms are matched with histograms of test images. Two sets of test images have been used, each containing 261 images of the same objects. The images of the two test–sets are taken under different conditions as the database images. The first test–set contains the following changes: 10° viewpoint changes for the Columbia image database, the second series of the aerial image of Marseille (including a viewpoint change) and between 10% and 15% scale changes for the remaining 61 objects.

measurement	$\cap$	$\cap_{we}$	$\cap_{mo}$	χ^2_v	χ^2_{qv}	χ^2_{av}	SSD	$maha$
recognition	98.9	82.4	47.9	95.4	99.2	97.3	96.9	98.1
errors	3	46	136	12	2	7	8	5

Table6.1: Recognition results on a database of 261 objects for the first test set. The test set includes 10% to 15% scale changes and $20°$ *viewpoint changes*

Table 6.1 shows the recognition rates for the first test–set obtained by 8 different histogram comparison measurements. The best recognition rate of 99% is obtained by the χ^2_{qv}–statistics. Nearly the same quality of results is given by the intersection measurement $\cap$. Recognition results above 95% are provided by the four measurements $maha$, χ^2_{av}, SSD and χ^2_v. The weighted intersection measurement $\cap_{we}$ still obtains a recognition rate of 82%, whereas $\cap_{mo}$ does not provide a satisfactory result.

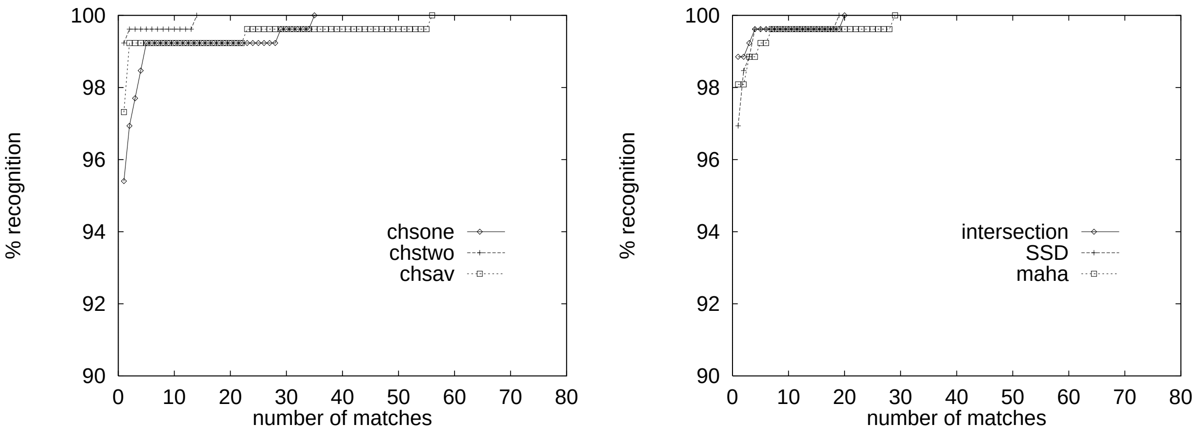


Figure6.1: Histogram matching for the first test set of 261 objects. The recognition rate is shown as a function of included matches

By taking into account not only the best match but the first n matches, we obtain the graph of figure 6.1. The horizontal axis shows the number of included matches for the calculation of the recognition rate. The recognition rate is shown by the vertical axis. In order to obtain a 100% recognition rate, χ^2_{qv} needs the first 14 matches. SSD and $\cap$ need about 20 matches and $maha$ 29 matches to attain 100% recognition. The number of matches for 100% recognition of these four measurements is in the order of 5% to 11% of the database size. χ^2_v needs 35 matches and χ^2_{av} already 56 matches. The remaining two weighted intersection measurements $\cap_{we}$ and $\cap_{mo}$ (not shown in figure 6.1) need both more than 100 matches.

The high recognition rates of χ^2_{qv} and $\cap$ in this first experiment show the applicability of histogram matching for object identification in the presence of changes in the appearance of objects. In particular for χ^2_{qv}, the correct match has been within the 5% best matches.

Table 6.2 shows recognition results of a second set of test images of the same 261 objects including more important changes: $20°$ viewpoint changes for 100 objects of the Columbia image database, the third series of the aerial images of Marseille (with more important viewpoint change) and 20% – 30% scale changes for the remaining 61 objects. Not surprisingly the recognition rates (for all comparison measurements) on this second test–set are less good than for

measurement	$\cap$	$\cap_{we}$	$\cap_{mo}$	χ^2_v	χ^2_{qv}	χ^2_{av}	SSD	$maha$
recognition	95.0	66.67	38.3	86.2	96.2	90.4	84.3	92.0
errors	13	87	161	36	10	25	41	21

Table6.2: Recognition results on a database of 261 objects for the second test set including 20% – 30% scale changes and $20°$ *viewpoint changes*

the first test–set. Table 6.2 shows a recognition rate of 96% for χ^2_{qv} indicating the robustness of this measurement to important changes in scale and viewpoint. The next best result is obtained by $\cap$ with 95% recognition. Above 90% recognition is obtained by χ^2_{av} and $maha$. Around 85% recognition is given by χ^2_v and SSD. As in the first experiment, the weighted intersection measurements $\cap_{we}$ and $\cap_{mo}$ provide worse results than all other measurements.

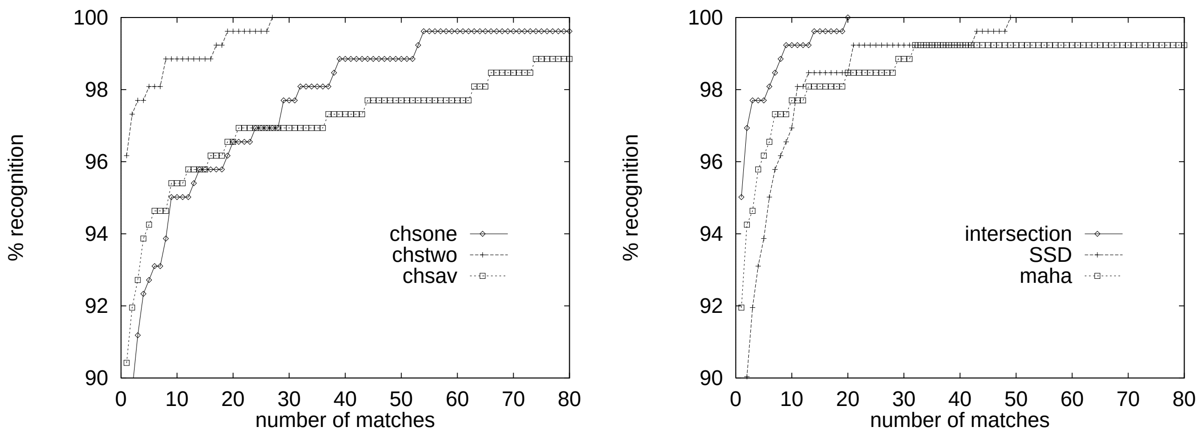


Figure6.2: Histogram matching for the second test set of 261 objects. The recognition rate is shown as a function of included matches

As for the first test–set, figure 6.2 shows the number of matches needed to obtain 100% recognition. The intersection measurement $\cap$ and χ^2_{qv} need the lowest number of matches. The number of matches is in both cases below 10% of the database size. SSD — with 49 matches — needs about 20% of the database size.

From these two initial experiments we can conclude that histogram matching provides a reliable means of object identification in the presence of important changes in the appearance of objects. The best comparison measurement has been χ^2_{qv}, followed by the intersection measurement $\cap$. Good results are obtained by the Mahalanobis distance $maha$, the sum of squared distances SSD, χ^2_{av} and χ^2_v. However, low recognition rates are provided by the weighted intersection measurements $\cap_{we}$ and $\cap_{mo}$. The following sections will show how to deal with appearance changes of objects in a more thorough manner.

6.2 Object identification in the presence of image plane rotation

The section describes experiments for the identification of objects in the presence of image plane rotation. As introduced in section 3.1.1 the response of a Gaussian filter at an arbitrary orientation can be calculated based on the responses of a finite set of Gaussian filters. By using this property of Gaussian derivatives — called steerability to image plane rotation — we calculate histograms of Gaussian derivatives of an arbitrary image plane rotation from a single image of the object. It is worth mentioning that not only Gaussian derivatives are steerable but for example also Gabor filters (see for details section 3.1.1).

In this section we apply four filter combinations: *Dx-Dy*, *Dx-Dy-Lap*, *Mag-Lap* and *Mag-G12-Lap*. The first two combinations are rotation variant and need to be steered to different image plane rotations. Applying steerability, we need only one image per object in order to calculate filter responses at arbitrary image plane rotations. Experiments have shown that the differences between steered rotations of $\Delta\alpha = 10°, 20°, 30°$ and $40°$ all provide very similar results for the identification of the 22 objects used. Since this result cannot be generalized to an arbitrary set of objects we do not report the analysis of $\Delta\alpha$ in detail. In the following we assume $\Delta\alpha = 20°$ and calculate the filter responses for 18 different rotations, namely for the angles $\alpha = 0°, 20°, 40°, \ldots, 340°$. For each of the 18 steered filter responses of an image we calculate a histogram. Each object is therefore represented by a collection of 18 histograms.

The remaining two filter combinations (*Mag-Lap* and *Mag-G12-Lap*) are invariant to image plane rotation. In contrast to the first two filter combinations we can use a single histogram for the representation of an object. Consequently, the number of histograms is significantly reduced. However, we anticipate a loss of rotational information, which may be significant for identification. Therefore, it will be particularly interesting to compare results of the rotation variant histograms to results of the rotation invariant histograms.

The used database of 22 objects is shown in figure A.2 of the appendix. For each object 18 different rotational views have been taken. The objects have been rotated manually in front of the camera so that the difference between each view is approximately 20°. The support for the histograms used throughout the section is circular in order to obtain the same portion of the object. The standard deviation σ of the Gaussian derivatives has been chosen to be $\sigma = 2.0$ and the radius of the circular support of each histogram has been $110\sigma = 220$ pixels. Each histogram is therefore calculated on the filter responses of about 150,000 pixel locations. The size of the images is 500×500 pixels. As a data set we have chosen randomly a single image per object. The test set contains all $22 \times 18 = 396$ images.

In a first experiment we examine the recognition results of the four filter combinations obtained by 8 different histogram comparison measurements (see section 5.1). The resolution of the histogram axes have been 32 so that the two–dimensional histograms (*Dx-Dy-32* and *Mag-Lap-32*) contain $32^2 = 1024$ cells and the three–dimensional histograms (*Dx-Dy-Lap-32* and *Mag-G12-Lap-32*) contain $32^3 = 32768$ cells. This number of cells is the maximal number of cells which have to be stored per histogram. As an analysis in section 6.7 shows, the number of occupied cells is much smaller and the histograms can be compressed.

Table 6.3 shows results obtained by the four filter combinations. The first table 6.3(a) shows the recognition results obtained by the best match between histograms of the test images and the database histograms. The best results (in most cases 100% recognition) have been obtained by the χ^2–statistics χ^2_{qv}. The intersection measurement $\cap$ provides nearly the same quality of results. Still very good results are obtained by the χ^2–statistics χ^2_{av} and the Mahalanobis distance *maha*. χ^2_v and SSD provide 90% recognition or higher in the majority of cases. Less than acceptable results are provided by the two weighted intersection measurements $\cap_{we}$ and $\cap_{mo}$.

measurement	$\cap$	$\cap_{we}$	$\cap_{mo}$	χ^2_v	χ^2_{qv}	χ^2_{av}	SSD	$maha$
Dx-Dy-32	99.5	29.0	65.7	95.7	100	97.5	88.9	99.0
Dx-Dy-Lap-32	100	56.6	16.7	97.5	100	98.7	95.5	99.7
Mag-Lap-32	99.8	43.4	30.6	91.9	99.5	99.5	96.7	92.7
Mag-G12-Lap-32	100	67.7	59.1	98.9	100	93.7	100	98.0

(a) recognition with first match

measurement	$\cap$	$\cap_{we}$	$\cap_{mo}$	χ^2_v	χ^2_{qv}	χ^2_{av}	SSD	$maha$
Dx-Dy-32	2	17	4	14	1	6	6	1
Dx-Dy-Lap-32	1	10	14	7	1	3	3	4
Mag-Lap-32	2	21	9	10	2	2	2	7
Mag-G12-Lap-32	1	18	8	7	1	4	1	2

(b) number of matches needed for 100% recognition

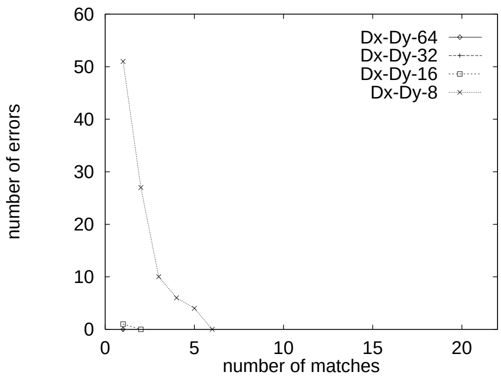
Table6.3: Recognition in the presence of image plane rotation

It is also interesting to examine the number of matches which is needed for correct recognition of all test images which is shown in table 6.3(b). χ^2_{qv} and $\cap$ need maximal 2 matches for recognition of all 396 test images. SSD, $maha$ and χ^2_{av} still need only a small number of matches. More matches are necessary for $\cap_{we}$, $\cap_{mo}$ and also for χ^2_v. Unfortunately, χ^2_v typically needs a high number of matches in order to obtain a 100% recognition rate even though good results are obtained with the first match.

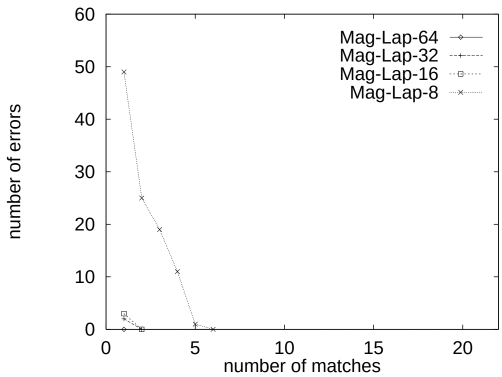
In table 6.3 we may compare two–dimensional histograms (*Dx-Dy* and *Mag-Lap*) with three–dimensional histograms (*Dx-Dy-Lap* and *Mag-G12-Lap*). Even though expected, we can confirm that the three–dimensional histograms give higher recognition results. This is explained by the additional information which is contained in the third independent histogram axis.

As mentioned above, it is also interesting to compare the rotation invariant histograms (*Mag-Lap* and *Mag-G12-Lap*) with those that are rotation variant (*Dx-Dy* and *Dx-Dy-Lap*). The results obtained are almost the same, independent of the rotational invariance. More important for recognition, is the use of the appropriate histogram comparison measurement and (as we will see below) the resolution of the histogram axes. Unfortunately, this is not always true: in the case of the aerial images of Marseille (see figures A.4 and A.5) rotation variant filter combinations provide much better recognition results than rotation invariant ones. In this case the additional storage requirement for the rotation variant histograms is justified. The choice between rotation invariant and rotation variant filter combinations depends strongly on the context and the database used. This choice may be based on the transinformation as motivated in section 4.2.3. The transinformation enables the evaluation of different measurement sets and therefore different filter combinations in the context of object recognition.

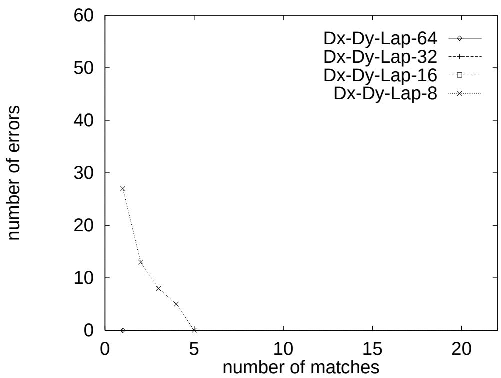
In order to examine the influence of the resolution per histogram axis we applied a number of different resolutions. For each of the four filter combinations we used resolutions of $8, 16, 32$ and 64 cells per histogram axis. Figure 6.3 shows the number of errors — for the same 396 test images as before — as a function of the number of matches. Here we used the χ^2–statistics χ^2_{qv} for histogram matching. As a first result we observe that for all four filter combinations the



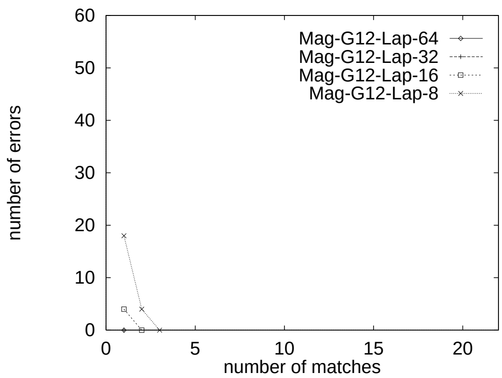
(a) *Dx-Dy*



(b) *Mag-Lap*



(c) *Dx-Dy-Lap*



(d) *Mag-G12-Lap*

Figure6.3: Recognition in the presence of image plane rotation for four filter combinations and four different resolutions

resolution of 64 is sufficient for a 100% recognition with the first match. Also a resolution of 32 is sufficient for three of the four filter combinations (except for *Mag-Lap*). The convergence for all filter combinations is fast. In the case of a resolution of 16 only the first two matches have to be considered in order to obtain 100% recognition. Furthermore, the convergence of the two two–dimensional filter combinations is very similar. The same statement is true for the two three–dimensional filter combinations which also converge in a similar way. Once again, the three–dimensional histograms provide better recognition results and faster convergence.

6.3 Object identification in the presence of scale changes

The section describes object identification experiments in the presence of scale changes. The experiments show that the approach can account for a scale factor in the order of 2. In order to manage such scale changes we use the equivariance property of Gaussian derivatives, which is described in 3.1.1. Larger scale changes may be considered by using an image pyramid.

In order to calculate histograms for responses of local descriptors at arbitrary scales we apply 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 3.1.1 and allows us to calculate the response of Gaussian derivatives at arbitrary scales. Given an image $g(x,y)$, the n^{th} order Gaussian derivative is defined by:

$$\frac{\partial^n}{\partial x^n} g(x,y) = G^{\sigma}_{x^n} \star g(x,y) \tag{6.99}$$

The n^{th} order derivative of the scaled image $f(x,y) = g(sx, sy)$ can be calculated on the basis of image $g(x,y)$ using the following equation (see 3.1.1 for details):

$$\frac{\partial^n}{\partial x^n} f(x,y) \quad = \quad s^n G^{\sigma s}_{x^n}(x,y) \star g(x,y) \tag{6.100}$$

Therefore, it is sufficient to adapt the calculation of Gaussian derivatives according to equation 6.100 in order to obtain the correct filter response for image $f(x,y)$ based on image $g(x,y)$ at a given image position (x,y) and at an arbitrary scale s. In order to calculate the histogram of Gaussian filter responses for a set of image positions, we need to adapt the positions (x,y) in the image $g(x,y)$. This can be done for example by the adaption 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. As we can see from equation $f(x,y) = g(sx,sy)$ the radius of the support region needs to be multiplied by s. This adaption of the support region is computationally inexpensive but compromises the precision (in particular for small $s < 1.0$). 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 to a predefined constant.

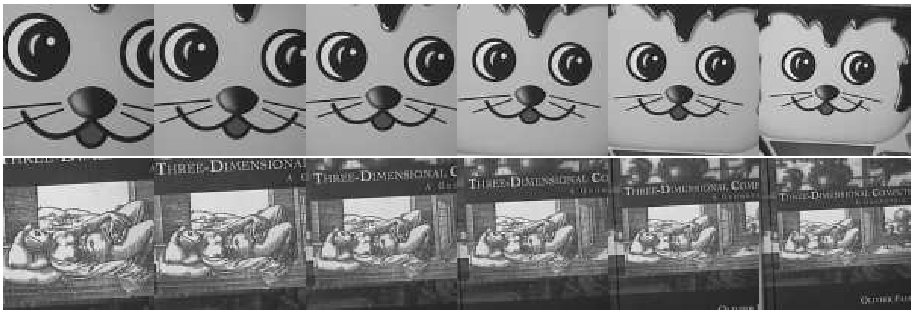


Figure6.4: The six scales for two objects of the database of 30 objects

The database used here contains 30 objects which are shown in figure A.1. For each object we recorded six images at six different scales. The scale factor between the first and the last image was approximately 2.1. The six images of the two objects are shown in figure 6.4. We have chosen one image per object (in figure 6.4 the third from the left) and we have calculated the two–dimensional histogram*Dx-Dy-64*: first derivative in the x– and y–directions and a resolution of 64 per histogram axis. The standard deviation of the Gaussian derivatives is $\sigma = 2.0$. The test set contains the remaining images corresponding to 5 different scales: $30 \times 5 = 150$ images. For each test image we calculated six histograms corresponding to scaled filter responses. The six scales were $\sigma = 1.23, 1.45, 1.7, 2.0, 2.35$ and 2.76, corresponding to 15% scale changes between each. The scale factor between the smallest and the largest scale is therefore 2.2. The radius of the support region for the histograms is 40σ and therefore varies between 49.2 and 94 pixels.

measurement	$\cap$	χ^2_v	χ^2_{qv}	χ^2_{av}	SSD	$maha$
recognition	99.3	96.7	100	98.0	92.0	98.0
number of matches for 100% recognition	2	5	1	2	12	5

Table6.4: Recognition results in the presence of scale changes for 30 objects. The results are for the two–dimensional histogram Dx-Dy-64

Table 6.4 shows the results of the experiment. The χ^2–statistics χ^2_{qv} obtains a 100% recognition rate for the 150 test images. The second best result is obtained by the intersection measurement $\cap$ with one error. This error is recognized with the second best match. Very good results obtain χ^2_{av} and $maha$ with 98% recognition. A slightly worse result is provided by χ^2_v. SSD does not obtain a satisfactory result in this experiment.

6.4 Object identification in the presence of scale changes and image plane rotation

The two previous sections 6.2 and 6.3 described experiments for object identification in the presence separately of image plane rotation and scale changes. This section describes two experiments for object identification in the presence of image plane rotation *and* scale changes. The first experiment uses the 22 objects of the image plane rotation experiment of section 6.2. The second experiment uses the database of both previous sections so that in all, the database contains 52 objects.

The first experiment uses the same 22 objects of the image plane rotation experiment (section 6.2, figure A.2). In addition to the 18 rotations of each of 22 objects we recorded two scaled images (per object and rotation). The scaling factor was approximately ± 1.75. For each of 22 objects and for each of 18 rotations per object the image database contains three different scales which are shown in figure 6.5. The overall number of images is therefore $22 \times 18 \times 3 = 1188$.

(a) scaling "out" (b) database image (c) scaling "in"

Figure6.5: The three scales of the first experiment of object identification in the presence of image plane rotation and scale changes. The scale changes are in the order of ± 1.75. *The overall scale factor between the left and right image is approximately 3.*

For the calculation of the histogram database we used one image per object (the middle scale of an arbitrary image plane rotation). Starting from this single image of an object we calculated $18 \times 3 = 54$ two–dimensional histograms corresponding to 18 different image plane rotations (respectively $\alpha = 0°, 20°, \ldots, 340°$) at three different scales. The scales have been set to $\sigma = 1.7, 3.0$ and 5.2. The histogram database therefore contains $22 \times 54 = 1188$ histograms. The used filter combination is *Dx-Dy-64* (first derivative in the x– and y–directions) with a resolution of 64 for each histogram axis.

As a test–set we used images of a different scale than those of the database: $22 \times 18 \times 2 = 792$ images. For each test image we calculate one *Dx-Dy-64* histogram with $\sigma = 3.0$. The results of the recognition experiment are shown in table 6.5. Table 6.5 shows that we obtain a 95.7% recognition with the first match (using χ^2_{qv}). By including the second best match we have a 100% recognition rate. Using $\cap$ for histogram matching we have a 91.2% recognition which reaches 100% is we include the first 6 matches.

measurement	$\cap$	χ^2_{qv}
recognition	91.2	95.7
number of matches for 100% recognition	6	2

Table6.5: Recognition experiment for 22 objects in the presence of scale changes and image plane rotation. Results are shown for the two–dimensional histogram Dx-Dy-64

It is interesting to note the difference between recognition of "scaling out" images and "scaling in" images. Since the considered scale changes are significant (scale factor of approximately ± 1.75), the recognition of the scaled images is difficult. Typically the "scaling in" images are recognized more reliably than the "scaling out" images. In this particular experiment no error has been obtained for the "scaling in" images. This can be explained by the smaller standard deviation σ which has to be used for the Gaussian derivatives. This increases the sensitivity to quantization noise and reduces the support region of the corresponding histograms. By using a larger σ we can reduce the sensitivity. On the other hand we lose details of the objects, which may be important for recognition.

measurement	$\cap$	χ^2_{qv}
recognition	98.1	99.1
number of matches for 100% recognition	8	2

Table6.6: Recognition experiment for 52 objects in the presence of scale changes and image plane rotation. Results are shown for the two–dimensional histogram Dx-Dy-64

Table 6.6 shows a second experiment for object identification in the presence of image plane rotation and scale changes. In this experiment we use all images of the experiments described in sections 6.2 and 6.3. The database therefore contains 52 objects with a scale factor of 2 for 30 objects and 18 image plane rotations for 22 objects. For calculation of the histogram database we use one image for each of the 52 objects. For each image we calculate histograms *Dx-Dy-64* corresponding to 18 image plane rotations $\alpha = 0°, 20°, 40°, \ldots, 340°$. Each object is therefore represented by 18 histograms. The histogram database contains $52 \times 18 = 936$ histograms.

As a test set we use 18 image plane rotations for 22 objects and 6 scaled versions of 30 objects. The number of test images is therefore $22 \times 18 + 30 \times 6 = 576$. For each test image we calculate six histograms *Dx-Dy-64* at different scales. As in section 6.3 we use the following standard deviations for the Gaussian derivatives: $\sigma = 1.23, 1.45, 1.7, 2.0, 2.35$ and 2.76. Table 6.6 shows that we obtain 99.1% recognition with χ^2_{qv}. By including the second best match the algorithm recognizes all test images correctly. Using the intersection measurement $\cap$ for histogram matching we obtain 98% recognition. 8 matches are necessary for 100% recognition.

6.5 Object identification in the presence of viewpoint changes

This section describes experiments made in order to evaluate the robustness of multidimensional receptive field histograms to viewpoint changes, using the Columbia image database [Mur 95]. The database contains 20 objects, each at 72 different viewing angles. Even though the database contains only viewpoints around a *viewing circle* rather than from a *viewing sphere* we are convinced that the results obtained here can be generalized to the more complicated case of a viewing sphere. In the experiments we examine the performance of different filter combinations, different histogram comparison measurements and design parameters of multidimensional receptive histograms.

Figure6.6: The 20 objects of the Columbia database

The Columbia database contains 20 objects (figure 6.6) seen from 72 different viewing angels, where each is $5°$ apart from the following (figure 6.7). Therefore the database contains $20 \times 72 = 1440$ images. Typically [Mur 95, Sch 96i] half of these images is used as a database (or learning set) and the remaining half as a test–set. In this case the angle between the viewpoints in the database is $\Delta\beta = 10°$.

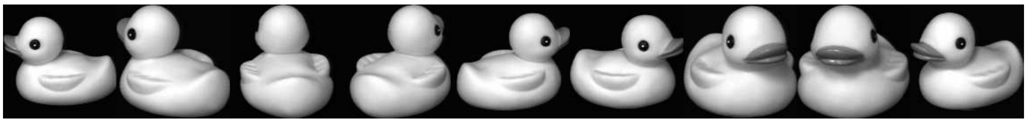


Figure6.7: 9 of 72 3D–rotations of one object of the Columbia database

The experiments described in this section demonstrate that histogram matching is relatively robust to viewpoint changes (3D–rotation). The experiments also show that we can use a low resolution for each axis of a receptive field histogram and still obtain high recognition rates. In particular, the section gives an indication of how many images of a 3D–object are necessary in order to recognize an object from an arbitrary viewpoint. As we know from the results of "aspect–graphs" the number of viewpoints in the database strongly depends upon the objects used. Therefore, the results obtained in this section should be seen as an indication and not as a general solution for modeling 3D–objects from 2D–images. We will discuss the problem more deeply in the context of object classification, where similar viewpoints of objects are clustered automatically (see chapter 9).

The experiments examine, in particular, the dependency of the robustness to 3D–rotation and the design parameters of histograms. The design parameters of histograms determine the separability between histograms of different objects (especially the number of local characteristics, which determine the number of dimensions of the histogram). An important parameter is the histogram resolution, or number of bins, which is used to represent each histogram dimension. Reducing the number of bins results in an improvement of the stability of the histogram, but also diminishes the discrimination between objects.

The section describes two different series of experiments: in the first series the relation between recognition results and histogram comparison measurements, filter combinations and the resolution of histograms is analyzed. A second series shows the dependency of recognition and the number of viewpoints (and therefore the angle between different viewpoints) in the image database.

Results of the first series of experiments

Table 6.7 shows the results of the first series of experiment where we use three different filter combinations: *Dx-Dy*, *Mag-Lap* and *Dx-Dy-Lap*. Each filter combination is calculated at different resolutions per histogram axis. The resolution varied between 2 cells and 64 cells per histogram axis. Eight different comparison measurements are used: three intersection measurements ($\cap$, $\cap_{we}$ and $\cap_{mo}$, see section 5.1.1), three χ^2–statistics (χ^2_v, χ^2_{qv} and χ^2_{av}, see section 5.1.3) and two quadratic form distances (SSD and $maha$, see section 5.1.2).

Table 6.7 shows 100% recognition by using for example the two–dimensional histogram *Dx-Dy-16* and the histogram comparison measurement χ^2_{qv}. 100% recognition is also provided by the three–dimensional histogram *Dx-Dy-Lap-8* using χ^2_{qv}.

filter	resolution	$\cap$	$\cap_{we}$	$\cap_{mo}$	χ^2_v	χ^2_{qv}	χ^2_{av}	SSD	$maha$
Dx-Dy	64	100	89.17	90.14	99.44	100	97.5	94.58	100
Dx-Dy	32	100	91.11	92.63	99.44	100	98.47	97.36	99.72
Dx-Dy	16	99.86	94.31	88.89	99.44	100	99.72	97.78	99.86
Dx-Dy	8	99.44	92.5	81.11	99.44	99.72	99.58	96.81	99.58
Dx-Dy	4	93.19	85.42	56.25	98.61	99.31	98.61	90.83	98.47
Dx-Dy	2	75.14	77.08	33.75	77.64	77.50	77.08	76.11	77.36
Mag-Lap	64	99.58	68.61	77.5	99.17	99.72	99.44	99.56	99.72
Mag-Lap	32	99.58	59.58	81.25	98.14	99.86	99.86	96.38	99.44
Mag-Lap	16	99.03	56.53	76.81	99.58	99.72	99.58	95.69	99.86
Mag-Lap	8	97.92	70.28	74.31	99.44	99.72	99.72	93.33	99.03
Mag-Lap	4	80.42	79.86	57.08	94.72	94.31	93.75	78.89	92.36
Mag-Lap	2	60.69	61.25	37.92	67.78	68.33	67.92	60.42	69.03
Dx-Dy-Lap	16	100	98.61	97.5	100	100	99.86	99.31	99.86
Dx-Dy-Lap	8	99.72	97.5	91.81	99.86	100	99.86	96.39	99.58
Dx-Dy-Lap	4	96.53	94.03	72.64	98.89	99.31	99.31	93.89	99.31
Dx-Dy-Lap	2	93.47	96.67	58.75	94.58	94.86	94.17	91.94	96.25

Table6.7: Recognition results of the first series of experiments

The results for *Dx-Dy* are slightly better than the results of *Mag-Lap*. This difference is due to the fact that *Dx-Dy* codes rotational information whereas *Mag-Lap* is invariant to image plane rotation. Table 6.7 shows better recognition results for the three dimensional histogram *Dx-Dy-Lap* relative to the two–dimensional histograms *Dx-Dy* and *Mag-Lap* in most cases. In particular in cases of low resolution the results of *Dx-Dy-Lap* are significantly better. For example a recognition rate of 96% is provided by *Dx-Dy-Lap-2* (using $maha$), where each histogram contains only $2^3 = 8$ cells. This gain in recognition can be explained by the fact that *Dx-Dy-Lap* contains three independent dimensions and therefore encodes more information than the two–dimensional histograms.

Table 6.7 also shows the results of eight different histogram comparison measurements. Overall the best results are obtained by the measurement χ^2_{qv}. The results of this measurement are better than the results of any other measurement, which shows its suitability to the matching of multidimensional histograms. The second best results (overall) are obtained by the Mahalanobis distance $maha$, χ^2_v and χ^2_{av}. The differences between these three measurements are relatively small. Overall the next best measurement is the unweighted intersection measurement $\cap$. Interestingly the intersection measurement obtains relative good results — with respect to χ^2_{qv} for example — when the resolution is high. That means that the results for $\cap$ are nearly as good as the results of χ^2_{qv} in the case of high resolution. In such cases the theoretical requirement of a sparse distribution (see section 5.1.1) for the application of $\cap$ is achievable. Unfortunately the results are significantly worse in the case of low resolutions as for example resolution 4. Still, relatively good results are obtained by using SSD for histogram matching. The two weighted intersection measurements $\cap_{we}$ and $\cap_{mo}$ give the worst results.

A last parameter, examined in these experiments, is the resolution per histogram axis. As we expect, the recognition rates increase with higher resolution. Since the stability decreases with a higher resolution there is a trade off between ability to discriminate and stability. As table 6.7 shows the increase of resolution until 16 always results in a gain in recognition rate. Increasing the resolution to 32 or even 64 decreases the recognition in some cases (for example for *Dx-Dy* using χ^2_{av} or SSD).

Robustness to viewpoint changes

In a second series of experiments we varied the angle between viewpoints stored in the image database. In particular, we used $\Delta\beta = 10°, 15°$, $20°$, $30°$, $40°$, $45°$, $60°$, $90°$. This enabled us to examine the robustness of histogram matching with respect to viewpoint changes. Besides $\Delta\beta$, we also varied the following parameters of the approach:

- the resolution of each histogram axis varied between 2 cells and 64 cells per histogram axis
- two filter combinations are used: *Dx-Dy-Lap* and *Dx-Dy*
- four histogram comparison measurements are applied: χ^2_{qv}, χ^2_{av}, $\cap$ and $maha$.

Figures 6.8 and 6.9 show recognition rates for different $\Delta\beta$, resolutions, histogram comparison measurements and filter combinations. The performance of the two–dimensional histogram *Dx-Dy* is less good than the performance of the three–dimensional histogram *Dx-Dy-Lap*. Even for large $\Delta\beta$ and low resolutions the filter combination *Dx-Dy-Lap* provides high recognition rates. This result indicates that high recognition rates can be obtained by adding dimensions to the histograms. The increase of memory can be partially compensated by a decrease in resolution (see section 6.7).

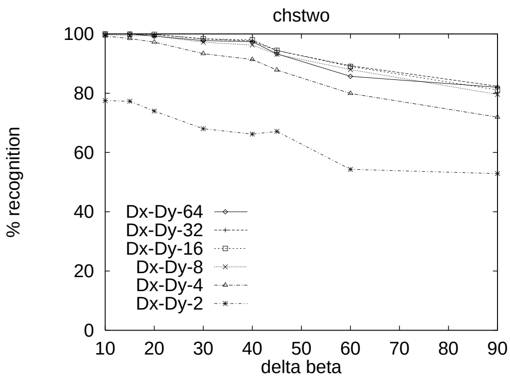
Figures 6.8 and 6.9 show recognition rates as a function of the angle $\Delta\beta$ between viewpoints of the database. Both filter combinations, all resolutions and all four histogram comparison measurements show a graceful degradation of the recognition rates with increasing $\Delta\beta$. This indicates the desired robustness of the approach to viewpoint changes. The slowest degradation is provided by χ^2_{qv} followed by the intersection measurement $\cap$. $maha$ gives slightly better results than χ^2_{av}.

6.6 Object identification in the presence of partial occlusion

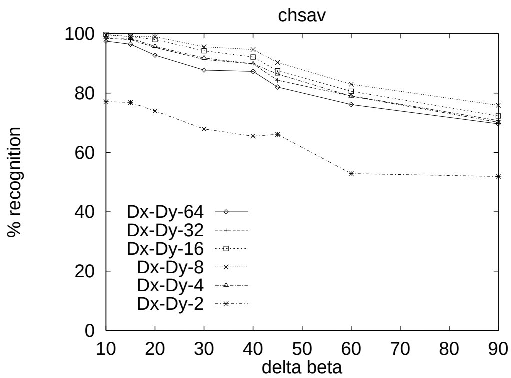
The section describes an experiment using 103 different objects. More precisely we use the following objects:

- 20 objects of the Columbia image database (figure 6.6)
- 30 objects of the scale experiment of section 6.3 (figure A.1)
- 22 objects of the image plane rotation experiment of section 6.2 (figure A.2)
- 31 additional objects (figure A.3)

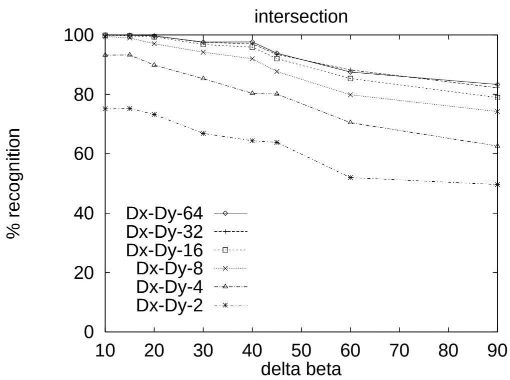
In the experiments described in previous sections we used two– and three–dimensional receptive field histograms. In the case of the 52 objects (section 6.4) we used for example two–dimensional histograms *Dx-Dy-64* which provided a recognition rate of 95.7%. As results of the



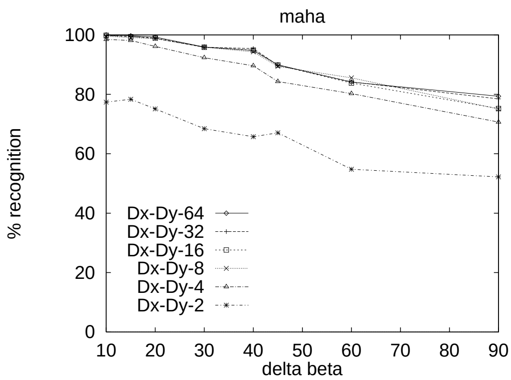
(a) χ^2_{qv}



(b) χ^2_{av}

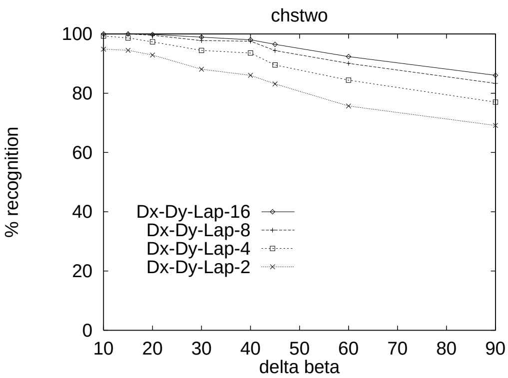


(c) $\cap$

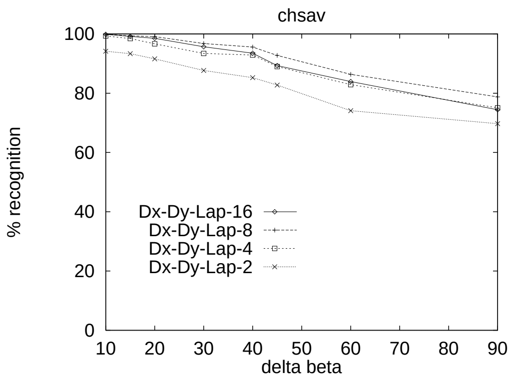


(d) *maha*

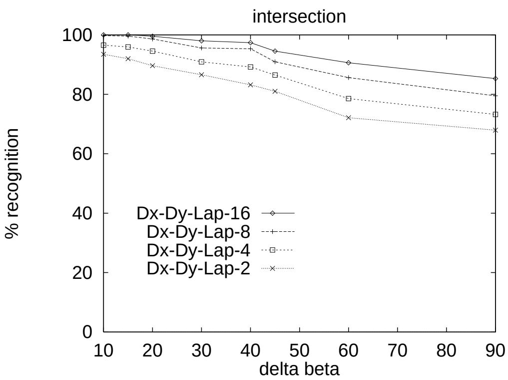
Figure6.8: Columbia database: two–dimensional histograms Dx-Dy. Relation between recognition rates and $\Delta\beta$ (see text).



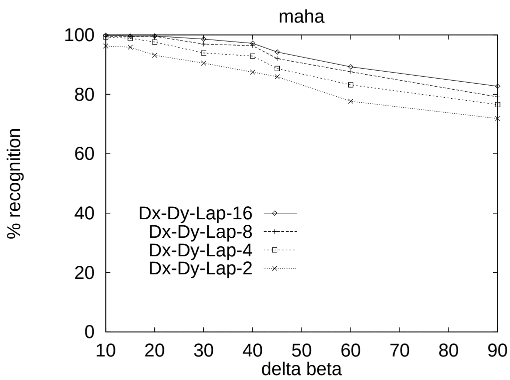
(a) χ^2_{qv}



(b) χ^2_{av}



(c) $\cap$



(d) *maha*

Figure6.9: Columbia database: three–dimensional histograms Dx-Dy-Lap. Relation between recognition rate and $\Delta\beta$ (see text).

previous sections indicate, recognition rates can be increased by adding independent dimensions to the histogram. Here we use a six–dimensional histogram of the filter combination *Dx-Dy-32* at three different scales with $\sigma_1 = \sigma$, $\sigma_2 = 2\sigma$ and $\sigma_3 = 4\sigma$. The results, reported below, show that these histograms contain sufficient information to recognize all of the test images of the 103 different objects.

For the Columbia image database we calculated $20 \times 36 = 720$ histograms corresponding to 36 different viewpoints of the 20 objects ($\sigma = 2.0$). Viewpoints being taken every $10°$. For the 30 objects used in the scale experiment and for the 22 objects used in the image plane experiment we calculated $52 \times 18 \times 6 = 5616$ histograms corresponding to 18 different image rotations and 6 different scales. We used image plane rotations of $\alpha = 0°, 20°, 40°, \ldots, 340°$. The different scales of the Gaussian derivatives were $\sigma = 1.48, 1.7, 2.0, 2.26, 2.62$ and 3.0. For each of the remaining 31 objects we calculated one histogram ($\sigma = 2.0$). The histogram database therefore contains $720 + 5616 + 31 = 6367$ histograms.

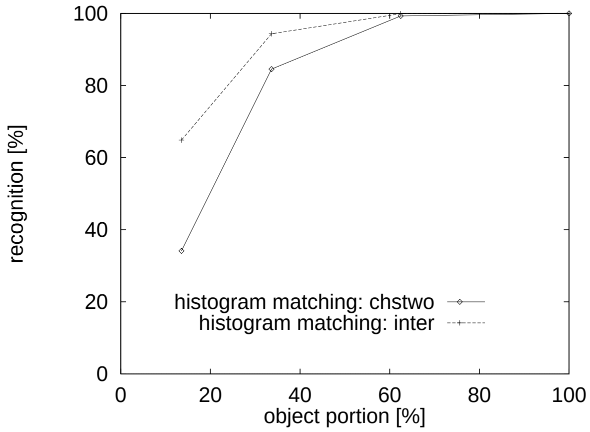


Figure6.10: Object identification in the presence of partial occlusion of 1327 images 103 different objects

The test set is composed of the following images: the remaining 720 images of the Columbia image database, 6 scaled images for each of the 30 objects of the scale experiment, 18 rotated images for each of the 22 objects of the image plane experiment and 31 different images for the additional 31 objects with slight changes with respect to the database images. The test set therefore contains $720 + 30 \times 6 + 22 \times 18 + 31 = 1327$ images. For each test image we calculated a single six–dimensional histogram with $\sigma = 2.0$.

Figure 6.10 shows the recognition results obtained by two different histogram comparison measurements: χ^2_{qv} and $\cap$. The recognition result is shown as a function of the visible portion of the objects. 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 6.10 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, $\cap$ still provides a recognition of 94%. χ^2_{qv} obtains 84% recognition in this case. The experiment emphasizes in particular the expected robustness of the intersection measurement $\cap$ with respect to partial occlusion.

Even though an object portion of about 62% is sufficient for the recognition of 103 objects the experiment corresponds to an ideal case: only a single object per image has been considered and the histograms have been centered on the object. In order to recognize objects in a cluttered scene we may calculate the histograms for a sliding window, corresponding to a certain portion of an object. In generally such an approach is time consuming and may even be infeasible for arbitrary shaped objects. Chapter 7 therefore develops an approach for the recognition of objects based only on a small number of image measurement vectors arbitrarily chosen in the image. The recognition results shown in figure 6.10 will serve as reference for the comparison of histogram matching and probabilistic object recognition proposed in chapter 7.

6.7 Memory requirement for multidimensional histograms

In the previous sections we concluded that an increase in the dimensionality of receptive field histograms typically results in better recognition. In the following, we shortly discuss an upper bound for the memory requirements of multidimensional receptive field histograms. Additionally we describe the memory requirement for two cases: firstly for 20 objects of the Columbia image database (see figure 6.6) and secondly for the experiment of section 6.6.

The maximal memory requirement of multidimensional receptive field histograms is given by the resolution R per histogram axis and by the number of dimensions L of the histogram: R^L. Since this maximal number is exponential in the number of dimensions, the storage requirement of multidimensional histograms needs further investigation.

An upper bound for the memory requirement is given by the size of the support region of a histogram. In the case of the 128^2 images of the Columbia image database the upper bound is for example in the order of $128^2 = 16384$. This is true simply because each single pixel position corresponds to a single measurement vector. Supposing that each measurement vector occurs exactly once in an image we obtain 16384 different measurement vectors which have to represented. Therefore the memory requirement of a histogram has an upper bound given by the image size.

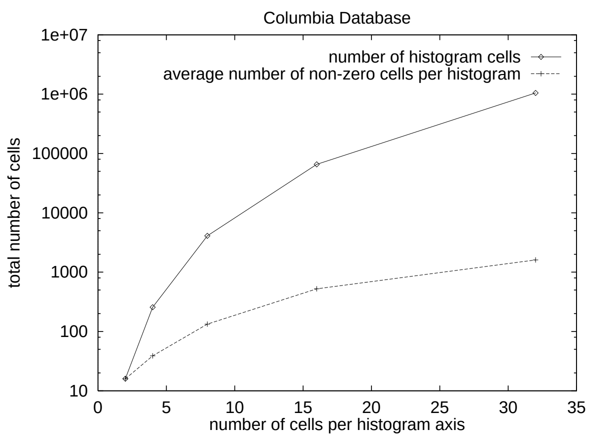


Figure6.11: Number of nonzero histogram cells for a four–dimensional histogram

In reality, measurements occur more than once in an image. The upper curve of figure 6.11

for example plots the theoretical number of histogram cells R^L. The lower curve of figure 6.11 shows the average number of different measurements (= number of nonzero histogram cells) as a function of the resolution R. The image database used here is the Columbia image database with $20 \times 72 = 1440$ images. The filter combination has been *Dx-Dy* at two different scales, namely for scale $\sigma = 2.0$ and 4.0. The corresponding histograms are therefore four–dimensional. Figure 6.11 shows that the average number of different histogram cells is much smaller than the theoretical size of the histograms. We use this fact to compress the histogram representation by a significant factor.

In section 4.1 we already analyzed the memory requirement for the experiment described in section 6.6. The average number of nonzero histogram cells was around 5000. Figure 4.2 shows the number of histogram cells as a function of their occurrence in the image. The six–dimensional histograms used in the experiment of section 6.6 are the most memory intensive histograms encountered in this thesis. Typically the average number of nonzero histogram cells was much smaller as shown for example in figure 6.11.

These two examples emphasize that the memory requirement for multidimensional receptive field histograms is in the order of several kB. When several image plane rotations, scaled versions and several viewpoints per object have to be stored, the memory requirement per object may be in the order of 1MB. Therefore algorithms for the automatic reduction of the representation are desirable. Such algorithms are proposed in chapter 9 in the context of object classification.

6.8 Conclusion

The experiments of this chapter show the applicability of histogram matching for object identification in the presence of scale changes, image plane rotation, viewpoint changes and partial occlusion.

From the experiments we can conclude the following:

- increasing the number of local characteristics and therefore increasing the dimensionality of receptive field histograms provides improved recognition results
- even low resolution histogram axes provide good recognition results. Nevertheless, increasing the resolution results in better recognition rates. A resolution of 32 has been shown to be sufficient in most cases
- the best overall histogram comparison measurement was χ^2_{qv}. This measurement typically provides good recognition results and also the lowest number of matches for 100% recognition.
- good recognition results are provided by the intersection measurement $\cap$. In particular in the presence of partial occlusion the intersection measurement $\cap$ performs better than χ^2_{qv}
- good recognition results are provided by χ^2_{av} and $maha$.
- histogram matching is robust to viewpoint changes, a graceful degradation of recognition results was reported in section 6.5
- the number of occupied histogram cells per histogram are in the order of 10^3 in the case of high–dimensional receptive field histograms.

In the case of partial occlusion the intersection measurement $\cap$ provided better results than χ^2_{qv}. A 100% recognition of 1327 images of 103 objects has been provided by $\cap$ using an object portion of 62% (see section 6.6). Nevertheless it is desirable to obtain higher robustness to partial occlusion. This motivates the development of the probabilistic object algorithm proposed in the following chapter 7. This algorithm is able to calculate the probability of an object only based on a small number of image measurement vectors arbitrarily chosen in the image.

Chapter 7

Probabilistic object recognition

From the preceding chapters we can conclude that multidimensional receptive field histograms provide a reliable means for the statistical representation of objects. In particular we analyzed different histogram *matching* techniques and showed their applicability for object identification in the presence of scale changes and image plane rotation as well as their robustness to viewpoint changes. As we applied the technique in the preceding chapters we did not explicitly consider *partial occlusion*. Nevertheless by using multidimensional histograms in a matching scheme we achieved a certain robustness to partial occlusion (see section 6.6).

In order to take partial occlusion into account in a thorough manner, this chapter develops a 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 any correspondence between the test-image and the object database.

The next section 7.1 develops a technique for the calculation of the probability of an object given a certain number of measurement vectors in the image. Sections 7.2 provides experimental results on a database of 52 objects in the presence of image plane rotation and scale changes. Section 7.3 uses the same database and the same test images as section 6.6. The 1327 test images of 103 different objects include scale changes, image plane rotation as well as viewpoint changes. We used the same six–dimensional histograms as in section 6.6 providing a basis for the comparison of *recognition by histogram matching* and *probabilistic object recognition* proposed in this chapter. In particular, 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. Section 7.4 proposes an extension of the algorithm for multiple object recognition in cluttered scenes.

7.1 Recognition without correspondence

In the context of probabilistic object recognition we are interested in the calculation of the probability of an object o_n given a local 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:

$$p(o_n|m_k) = \frac{p(m_k|o_n)p(o_n)}{p(m_k)} = \frac{p(m_k|o_n)p(o_n)}{\sum_i p(m_k|o_i)p(o_i)} \tag{7.101}$$

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:

$$p(o_n|m_k \wedge m_j) = \frac{p(m_k \wedge m_j|o_n)p(o_n)}{\sum_i p(m_k \wedge m_j|o_i)p(o_i)} \tag{7.102}$$

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

$$p(o_n|m_k \wedge m_j) = \frac{p(m_k|o_n)p(m_j|o_n)p(o_n)}{\sum_i p(m_k|o_i)p(m_j|o_i)p(o_i)} \tag{7.103}$$

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

$$\begin{aligned} p(o_n|\bigwedge_k m_k) &= \frac{p(\bigwedge_k m_k|o_n)p(o_n)}{\sum_i p(\bigwedge_k m_k|o_i)p(o_i)} && (7.104) \\ &= \frac{\prod_k p(m_k|o_n)p(o_n)}{\sum_i \prod_k p(m_k|o_i)p(o_i)} && (7.105) \end{aligned}$$

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 7.105 relies on the fact that all K measurement vectors come from the same object. Experiments will show that already a small number K of measurement vectors and therefore a small visible portion of an object provide reliable object hypothesis. This implies that we can in general safely assume that the K measurement vectors are in fact from a single object. Obviously a rough figure–ground segmentation would help whenever possible but the results indicate that the algorithm can be applied without such pre–processing.

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 experiments of this chapter (sections 7.2 and 7.3) and later chapters we therefore 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 7.105 simplifies to:

$$p(o_n | \bigwedge_k m_k) = \frac{\prod_k p(m_k|o_n)}{\sum_i \prod_k p(m_k|o_i)} \tag{7.106}$$

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 7.106 shows a calculation of the probability for each object o_n entirely based on the multidimensional receptive field histograms of N objects.

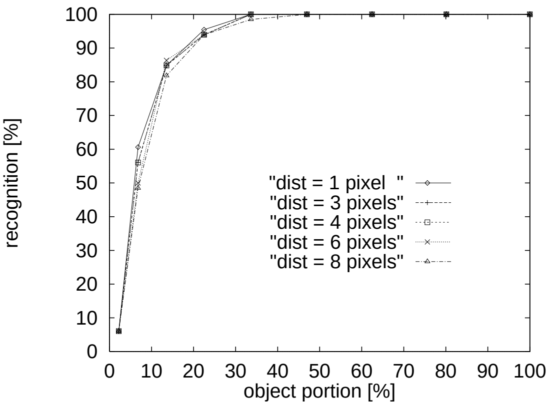
It is important to note that the locations of the measurement vectors can be chosen arbitrarily. Consequently the technique is fast (only a certain number of local receptive field vectors have to be calculated) and robust to partial occlusion (the approach is strictly local). Furthermore, the technique works without correspondence between the object database and the test image.

The next two sections describe experiments of the application of equation 7.106: the next section 7.2 describes an experiment with 52 objects in the presence of image plane rotation and scale changes. The experiment in the following section 7.3 increases the number of objects to 103 and additionally considers viewpoint changes for 20 objects.

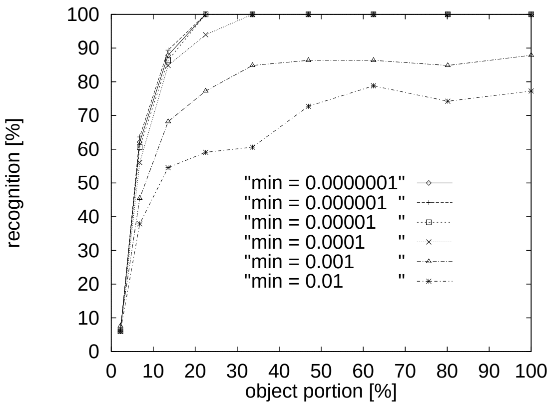
Experimental evaluation of implementation parameters

An important assumption of equation 7.106 is the independence of the K different local measurement vectors m_k. In order to satisfy the independence requirement, a certain distance $d(m_k, m_j)$ between different measurement vectors has to be taken into account. From a signal processing point of view and because of the Gaussian envelope of the Gaussian derivatives, a distance $d(m_k, m_j) \geq 3\sigma$ is sufficient for mutual independence between any pair of measurement vectors m_k and m_j. A distance of $d(m_k, m_j) \geq 2\sigma$ corresponds to weak dependency between measurement vectors.

In the context of probabilistic object recognition we are principally interested in the calculation of probabilities using equation 7.106 for a local image region. It is desirable that this image region be as small as possible. In this context, two influences create a tradeoff: on the one hand the assumption of independence demands a certain distance $d(m_k, m_j)$ but on the other hand increasing this distance decreases the number K of measurement vectors for a chosen local image region. We evaluated experimentally the influence of distance $d(m_k, m_j)$ for fixed image regions.



(a) *different distances $d(m_k, m_j)$, p_{min} = 0.0001 constant*



(b) *different minimal probabilities p_{min}, $d(m_k, m_j)$ = 4 pixels constant*

Figure7.1: Empirical evaluation of two parameters: (a) $d(m_k, m_j)$. (b) p_{min}. Database of 22 objects and 66 test images

Figure 7.1(a) shows the influence of distance $d(m_k, m_j)$ on the recognition of 22 objects. For each of the objects we calculated six–dimensional histograms corresponding to 18 different rotations and 6 different scales (see for details section 7.2). The histogram database therefore contains $22 \times 18 \times 6 = 2376$ histograms. Three rotated images for each of 22 objects have been chosen arbitrarily as test set. Figure 7.1(a) shows the recognition rate as function of the visible portion of the objects in the image. Different graphs are obtained for different distances $d(m_k, m_j)$ between measurement vectors. As mentioned above, the total number K of measurement vectors for a fixed object portion increases according to the decrease of distance $d(m_k, m_j)$. Results are shown for distances $d(m_k, m_j)$ of $1, 3, 4, 6$ and 8 pixels between measurement vectors. Surprisingly the recognition results for all distances are nearly identical. This result can be explained by the compensatory effect, the tradeoff between the influence of independence and number K of different measurement vectors. Only the largest distance of 8 pixels provides slightly worse results. In the experiments below we therefore set the minimal distance $d(m_k, m_j)$ between two measurement vectors m_k and m_j to 4 pixels (corresponding to 2σ).

The influence of another parameter is shown in figure 7.1(b): the minimal probability $p_{min} \leq p(m_k|o_n)$ which is used in equation 7.106. Such a minimal probability is necessary since the estimation of the probabilities $p(m_k|o_n)$ is sensitive to signal disturbances for small $p(m_k|o_n)$. Such disturbances can be caused, for example, by changes in the appearance of objects. In order to be robust to such changes we associate a minimal probability p_{min} to all multidimensional histograms. Figure 7.1(b) shows the recognition rates as a function of the visible object portion obtained for the same 66 test images of 22 objects as before. Different graphs for p_{min} between 10^{-2} and 10^{-7} are shown. The graphs motivate the choice of $p_{min} = 10^{-5}$ since no further improvement of recognition is obtained by smaller p_{min} and smaller values would increase the sensitivity to signal disturbances (not examined in the figure).

7.2 Object identification in the presence of scale changes and image plane rotation

The section describes an experiment for object identification for 52 objects in the presence of combined scale changes and image plane rotation. The experiment described uses the same database and the same test images as the second experiment of section 6.4. 22 of the objects are shown in figure A.2 and the remaining 30 objects are shown in figure A.1. Instead of the histogram matching techniques used in section 6.4 this section applies equation 7.106 for probabilistic object recognition.

This and the following section use the filter combination *Dx-Dy* at three different scales: $\sigma_1 = 2.0, \sigma_2 = 4.0$ and $\sigma_3 = 8.0$. Each measurement vector m_k and the corresponding histograms are six–dimensional. As in section 6.4 we employ the steerability of Gaussian derivatives in order to calculate filter responses at arbitrary image plane rotations. We thus steer the filter combination to 18 different orientations with a distance between each orientation of $\Delta\alpha = 20°$.

By employing the equivariance property of Gaussian derivatives it is possible to calculate filter responses at arbitrary scales from a single image of the object. Since the test images of the objects cover a scale change of approximately factor 2.2, we calculate six scaled versions for the database images of each object (with the same σ as in section 6.4: $\sigma_1 = 1.23, 1.45, 1.7, 2.0, 2.35$ and 2.76, and correspondingly $\sigma_2 = 2\sigma_1$ and $\sigma_3 = 4\sigma_1$).

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 below). 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 49 pixels (for $\sigma_1 = 1.23$) and 110 pixels (for $\sigma_1 = 2.76$). Therefore the support region of the histograms differs up to a factor of $5 \approx \frac{110^2}{49^2}$. The centering of the support region can be seen as a figure–ground segmentation for learning an object model.

The histogram database is calculated on the basis of a single image per object and contains $52 \times 18 \times 6 = 5616$ histograms corresponding to 18 different rotations and 6 different scales for each of 52 objects.

In order to apply the probabilistic object recognition algorithm (equation 7.106), K independent measurement vectors m_k have to be chosen. As mentioned above, two assumptions underlying equation 7.106 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 $2\sigma_1$. 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. Section 7.4 defines an extension to the algorithm that recognizes multiple objects in cluttered scenes.

The set of test images is composed of 18 rotated versions of the 22 objects of figure A.2 and 6 different scales of the 30 objects shown in figure A.1. The test set therefore contains $22 \times 18 + 30 \times 6 = 576$ images. As introduced above, we calculate measurement vectors for each test image lying within a square region at the center of the image. The radius (half side length of the square) has been varied from $1\sigma_1, 5\sigma_1, 10\sigma_1, 15\sigma_1, \ldots, 50\sigma_1$ pixels, corresponding to 1, 25, 100, 225, ..., 2500 measurement vectors. The maximal radius of $50\sigma_1$ covers a larger region than the support region of the database histograms (which have been calculated with radius $40\sigma_1$). Therefore we can report the influence of background – which is not modeled explicitly – on the recognition results. The six–dimensional filter combination *Dx-Dy* at three scales is calculated with $\sigma_1 = 2.0, \sigma_2 = 4.0$ and $\sigma_3 = 8.0$.

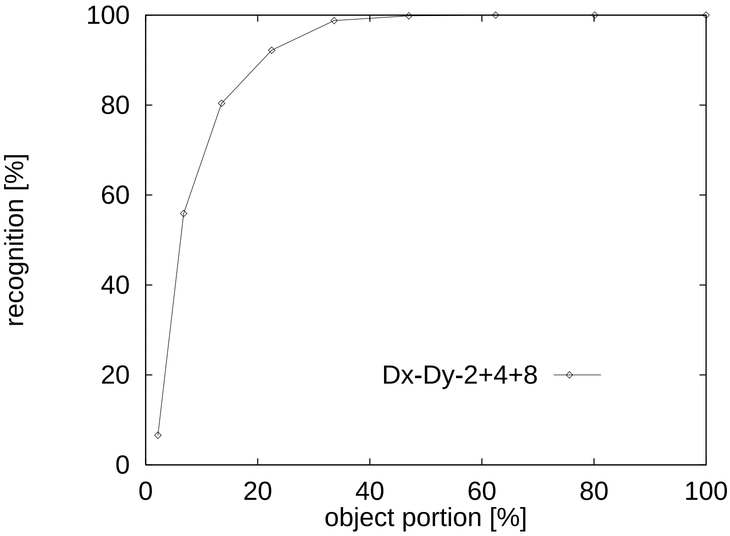


Figure7.2: Experimental results with 52 objects

Figure 7.2 and table 7.1 summarize the recognition results for the 576 test images. The first result is that approximately 62% visibility of the objects is sufficient for the recognition of all test images. With about 34% visibility of the objects only 7 errors occur corresponding to a recognition rate of nearly 99%. By using 13.5% of the object (more than 85% occlusion) more than 80% of the test images are recognized.

radius $[\sigma_1]$	1	5	10	15	20	25	30	35	40	45	50
object portion [%]	2.2	6.8	13.5	22.5	33.6	47.0	62.5	80.1	100.0	122.1	146.3
recognition [%]	6.7	56.0	80.4	92.2	98.8	99.8	100	100	100	100	100
errors	538	254	113	45	7	1	0	0	0	0	0

Table7.1: Experimental results with 52 objects

Table 7.1 shows 100% recognition also for the radii 45 σ_1 and 50 σ_1. These radii correspond to 122% and 146% of the support region of the database histograms. For these image regions the influence of the background is important. The 100% recognition in these cases indicates that the algorithm is able to recognize all objects correctly even without a good figure ground segmentation and without an explicit model of the background.

Overall, the experimental results show that the proposed approach can deal with significant partial occlusions. Furthermore, a small object portion is sufficient in order to obtain a good object hypothesis.

7.3 Object identification in the presence of partial occlusion

The previous section described an object identification experiment for 52 objects in the presence of image plane rotation and scale changes. In order to consider viewpoint changes this section adds 20 objects from the Columbia image database (see figure 6.6). For each of 20 objects we add 36 viewpoints where each viewpoint is 10° apart. The 31 objects, shown in figure A.3 are also considered. The entire object database contains 103 different objects and corresponds to the database used in section 6.6.

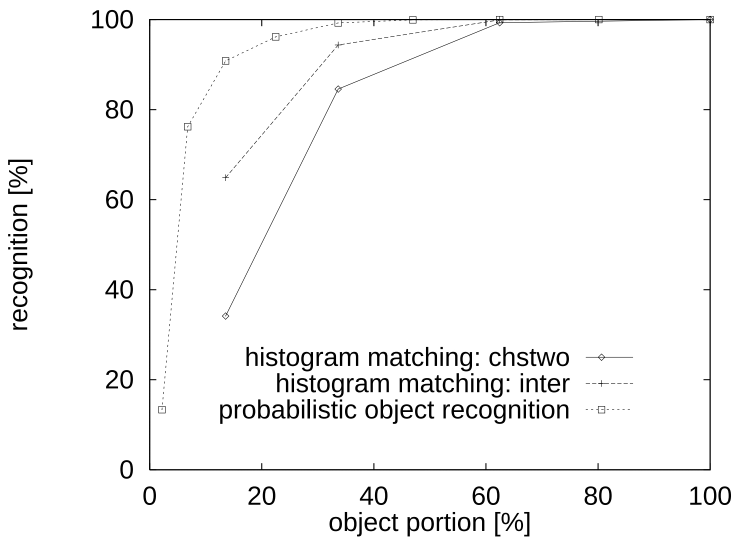


Figure7.3: Experimental results for 103 objects. Comparison of probabilistic object recognition and recognition by histogram matching: χ^2_{qv} (chstwo) and $\cap$ (inter)

The technical details (filter combination, σ_i, radius of histogram support region) of the experiment described here are the same as in the previous section 7.2. So we calculate $52 \times 18 \times 6 = 5616$ histograms for the 52 objects used before. For the 36 viewpoints of the 20 objects of the Columbia image database we calculate $20 \times 36 = 720$ histograms. For the remaining 31 objects we add one histogram per object. Overall the histogram database contains $5616 + 720 + 31 = 6367$ histograms.

As test set we use 576 images of the 52 objects (as in the previous section). For the Columbia image database we use the remaining 36 viewpoints per object, which adds an additional 720 images to the test set. For the remaining 31 objects we use test images, where slight changes (light intensity, rotation and scale) are introduced for each object. The test set contains 1327 images and corresponds to the test set used in section 6.6.

Figure 7.3 and table 7.2 summarize the recognition results. Even though the number of different objects is nearly doubled relative to the previous section, the recognition results are nearly the same. Due to the simplicity of the Columbia image database the recognition *percentages* are even better than before. A visible object portion of approximately 62% is sufficient for the recognition of all 1327 test images (the same result has been provided by histogram matching in section 6.6). With 33.6% visibility the recognition rate is still above 99% (10 errors in total). Using 13.5% of the object the recognition rate is still above 90%. More remarkably, 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.

radius $[\sigma_1]$	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
recognition [%]	13.3	76.2	90.8	96.2	99.3	99.9	100	100	100
errors for the 52 objects	547	255	115	48	9	1	0	0	0
errors for Columbia database	573	42	0	0	0	0	0	0	0
errors for last 31 objects	30	19	7	3	1	0	0	0	0

Table7.2: Experimental results with 103 objects

Section 6.6 uses exactly the same six–dimensional histograms, the same histogram database and the same 1327 test images. Figure 7.3 therefore shows the recognition results obtained by the probabilistic object recognition algorithm and by histogram matching. In particular, results for the χ^2–statistics χ^2_{qv} and the intersection measurement $\cap$ of section 6.6 are shown. As we can see the robustness to partial occlusion is significantly increased by applying the probabilistic object recognition scheme developed in this chapter.

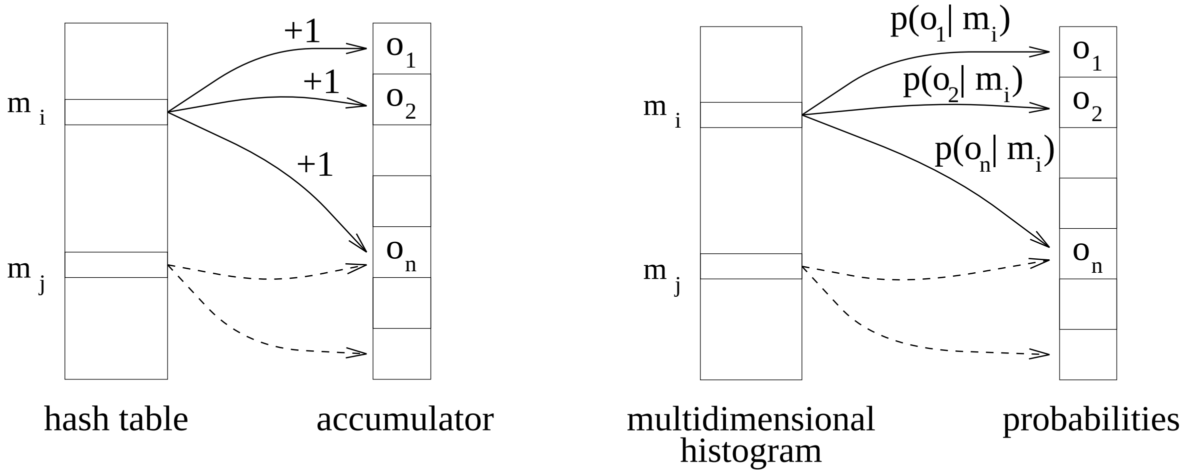
We can conclude that the proposed probabilistic object recognition approach is capable to discriminate 103 objects in the presence of significant scale changes, image plane rotation and viewpoint changes. Furthermore, the approach is robust with respect to partial occlusion since a small portion of the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

7.4 Probabilistic object recognition in cluttered scenes

Section 7.1 defined a probabilistic object recognition algorithm calculating the probability $p(o_n|R) = p(o_n| \wedge_k m_k)$ for all objects o_n based on a local image region R. The experiments in sections 7.2 and 7.3 showed that the algorithm is capable to recognize 103 objects in the presence of image plane rotation, scale changes and viewpoint changes. Nevertheless, experimental results have been shown only for the case of a single object per image. This section discusses an extension of the algorithm in the context of multiple object recognition in cluttered scenes.

The results of the preceding sections indicate that we can define an algorithm for object recognition in the context of cluttered scenes. Such an algorithm can, for example, employ a hypothesis–testing strategy as described in chapter 8. Another possibility is to use local image regions for probabilistic voting for objects. Such an algorithm resembles object recognition over a hashtable, where image regions correspond to features of the object. The main advantage of hashing based on our algorithm is that the size and form of image regions (= features) can be chosen dynamically and even at *runtime*. In the following we define such an algorithm as an extension of our probabilistic object recognition algorithm of section 7.1.

A standard algorithm for recognition in cluttered scenes is based on a hashtable and an accumulator for each object. Schmid and Mohr have recently shown the applicability of such an algorithm based on discriminant vectors of local characteristics (see section 2.2.3). Since our algorithm defined in section 7.1 has structural similarities to a hashtable based algorithm we will discuss both approaches and define an extension of our probabilistic object recognition algorithm synthesizing both approaches.



(a) *Hashtable based algorithm*

(b) *Probabilistic object recognition*

Figure7.4: Structural outline of recognition over a hashtable and probabilistic object recognition

Figure 7.4(a) shows the general scheme of a hashtable based algorithm. The algorithm calculates indices m_k from a test image. Using one index m_i as index of the hashtable we obtain a set of objects, which could have produced the index m_i. Typically, the accumulator of each object of the set is increased by one (or an appropriate certainty factor). This step of the algorithm is called voting. The algorithm repeats the hashing for indices m_k until an object has accumulated a sufficient number of votes.

Figure 7.4(b) shows the scheme of our probabilistic object recognition algorithm. The algorithm uses an index m_i in order to calculate probabilities $p(o_n|m_i)$ (using equation 7.101) for all objects o_n. The algorithm repeats the calculation of indices m_k and incorporates the probabilities $p(o_n|m_k)$ into the calculation of probabilities $p(o_n|R) = p(o_n| \wedge_k m_k)$ (see equation 7.106) of all objects and a local image region R. The algorithm may stop as soon an object has a sufficiently high probability.

Even though the structure of both algorithm is similar, significant differences exist between them. We discuss their advantages and disadvantages based on the following criteria:

- "hard" decision against "soft" decisions
- consistency test of information coming from different image locations
- underlying features

Since a hashtable based algorithm accumulates the votes for different objects in an accumulator array the decision is "hard" in the sense that it is decided if an object did or did not produce an index. Since the voting for different indices is independent, all possible combinations of indices are considered. This typically results in false positives which motivate the application of consistency tests between different indices. Even though such consistency tests typically decrease the false positive rate, they have to be coded into the hashtable prior to recognition.

The probabilistic recognition algorithm defined in section 7.1 uses soft decisions by computing probabilities for each object. One of the major differences to hashtable algorithms derives from the fact that equation 7.106 supposes that all indices m_k come from the same object. Therefore the consistency test is done for each single object separately, without considering any combination of different objects. This makes the algorithm inappropriate in the context of cluttered scenes with multiple objects. Since the hashtable enables such a consistency test for multiple objects in the scene we want to define an extension of our algorithm based on properties of both approaches.

A last interesting criteria for the purpose of comparison are the underlying features. On the one hand the features in a hashtable based approach are fixed and chosen prior to application. On the other hand in our approach, the size and form of an image region R used for the calculation of the probabilities $p(o_n|R)$ may be chosen dynamically at runtime and without recalculation of the multidimensional histograms. We can interpret an image region R as dynamical adaptable index of a hashtable.

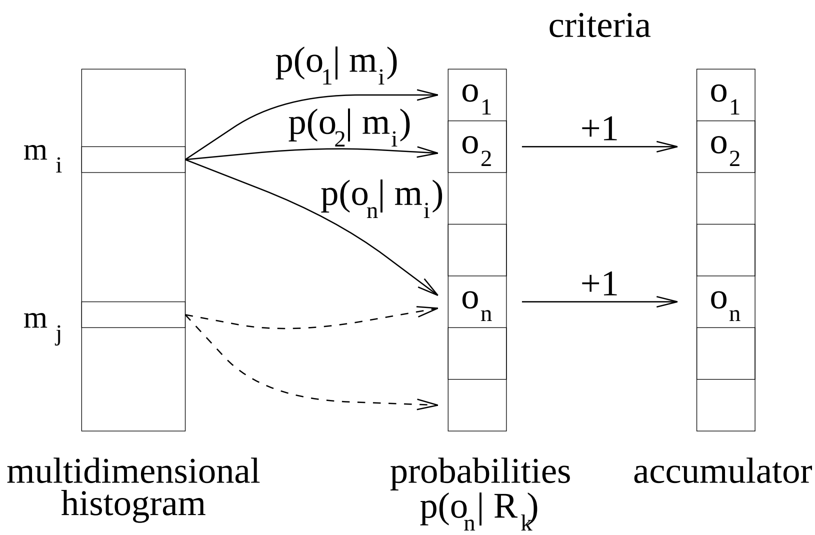


Figure7.5: Probabilistic object recognition in cluttered scenes

Figure 7.5 presents the idea of object recognition over a hashtable indexed by dynamically adaptable image regions R_k. A "dynamic" hashtable indexed by image regions R_k is implemented by our algorithm (in the figure shown as multidimensional histogram and probability array $p(o_n|R_k)$). As soon as certain criteria are fulfilled (for instance the probability is above a threshold or the image region has reached a predefined size) the algorithm votes for a set of objects and increases the accumulator of these objects. In order to apply this algorithm, appropriate *criteria* have to be defined. In principal, two types of criteria may be used: criteria concerning the local image regions R_k (for example size, amount of discriminant information, consistency of region) and/or probability $p(o_n|R_k)$ criteria (for example above a certain threshold). In either case the criteria can be changed dynamically during runtime of the algorithm without recalculation of the "hashtable indexed by image regions R_k".

In order to illustrate the proposed recognition algorithm based on a "dynamically indexed hashtable", we describe an experiment on a database of 100 objects. The task is to recognize objects in cluttered scenes as shown in figures 7.6(a), (c) and (e). For each of 100 objects we have recorded one image (256 × 256 pixels) and computed the six–dimensional histogram *Mag-Lap-32* (Magnitude of first derivative and Laplacian operator, resolution of 32 cells per histogram axes) at three different scales, namely $\sigma_1 = 2.0, \sigma_2 = 4.0$ and $\sigma_3 = 8.0$. The image region R_k has 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 becomes the score in the accumulator. Ultimately, the highest "scores" of the accumulator array correspond to objects in the scene.

Figure 7.6(a) shows an example scene and figure 7.6(b) shows the four objects with highest scores of the accumulator array for all 100 objects. We can see that the four objects in the scene correspond to the first four candidates. Another testimage is shown in figure 7.6(c) and the four best candidates are given by figure 7.6(d). Once again all four candidates are contained in the scene. A last example of a cluttered scene is given in figure 7.6(e). Three objects are visible in the foreground. These three objects are contained within the first three matches shown by figure 7.6(f). Interestingly the fourth match shows another part of the object of the first match. This emphasizes that different objects which are similar (here "simulated" by different parts of the same object, but having overlapping parts) compete with each other and ultimately the correct object is recognized.

Overall we can conclude that the proposed algorithm combines advantages of our probabilistic object recognition algorithm with advantages of a hashtable based approach. By adapting the image regions R_k and by combining the results appropriately, we may obtain a powerful object recognition algorithm strategy, which can be applied in many circumstances. The results indicate that we may apply our probabilistic object recognition algorithm for cluttered scenes by employing a "dynamically indexed" hashtable.

7.5 Conclusion

In this chapter we have extended the *matching* of multidimensional receptive field histogram to *probabilistic object recognition without correspondence*. We have developed a technique to determine the probability that an object may be found in an image based only on multidimensional receptive field histograms. Promising results have been shown on a database of 103 objects in the presence of scale changes, image-plane rotation and viewpoint changes.

(a) *Test image 1*

(b) *First matches for test image 1*

(c) *Test image 2*

(d) *First matches for test image 2*

(e) *Test image 3*

(f) *First matches for test image 3*

Figure7.6: Experimental results for three test images with cluttered scenes

In the last section we have extended probabilistic object recognition to a "dynamically indexed hashtable". The algorithm allows the recognition of multiple objects in cluttered scenes by incorporating the advantages of a hashtable based approach. As "indices" we use image patches and apply our probabilistic object recognition algorithm in order to obtain object sets which may correspond to the image patch. The main advantage of the proposed approach is that the image patches can be adapted dynamically at runtime with respect to context dependent criteria. Promising experimental results are shown on a database of 100 objects.

Chapter 8

Active object recognition

Chapter 7 extended the matching of multidimensional receptive field histograms to a probabilistic object recognition algorithm. The main result of the experiments was that the algorithm provided a reliable object hypothesis from a small local image region. This allowed us to define, in section 7.4, an extension of the algorithm which may recognize multiple objects in cluttered scenes by using a dynamically indexed hashtable.

This chapter discusses another possible extension of the probabilistic object recognition technique, introducing an *active object recognition* framework. This framework is for the most part based on reliable object hypothesis and estimates of the object's pose provided by the probabilistic object recognition algorithm. A pose estimate is obtained since the recognition algorithm calculates the probability for a collection of multidimensional receptive field histograms with each corresponding to a particular object o_n and a particular pose of the object S_j including scale t_z and rotation $(r_x r_y r_z)$ of the object (see section 4.1).

This chapter discusses the possibilities for active object recognition in 2D and 3D as an extension of probabilistic object recognition as defined in section 7.1. The general framework is given by:

1. Hypothesize object and object pose (by using histogram matching or probabilistic object recognition):

$$\begin{aligned} \hat{o}_n(t-1) &\equiv \text{hypothesized object at time } t-1 \\ \hat{S}_j(t-1) &\equiv \text{hypothesized pose at time } t-1 \end{aligned}$$

2. Move camera to pre–computed salient region in 2D or salient viewpoint in 3D:

$$\Delta S(t-1) \quad \equiv \quad \text{parameter of camera movement}$$

3. Verification: difference between expectation and observation:

$$\begin{aligned} \hat{o}_n(t) &= \hat{o}_n(t-1) \\ \hat{S}_j(t) &= \hat{S}_j(t-1) + \Delta S(t-1) \pm \epsilon \end{aligned}$$

The hypothesis of an object and its pose may be given by the maximal probability $p(o_n|R, S_j)$ (equation 7.106) calculated for a local image region R and all poses S_j of the object:

$$(\hat{o}_n, \hat{S}_j) : \max_{n,j} p(o_n|R, S_j) \tag{8.107}$$

The multidimensional histogram $h(M|o_n, S_j) = h(M|o_n, t_z, r_x, r_y, r_z)$ corresponding to the maximal probability provides us with an estimate of the object $\hat{o}_n$ and its pose $\hat{S}_j = (t_z, r_x, r_y, r_z)$.

Alternatively, the hypothesis of an object and its pose may be given by the minimal χ^2–statistics calculated for an image I and its histogram $H(M|I)$:

$$(\hat{o}_n, \hat{S}_j) : \min_{n,j} \chi^2_{qv}(H(M|I), H(M|o_n, S_j)) \tag{8.108}$$

In order to verify the object hypothesis we want to move the camera to a preselected discriminant position (in 2D or 3D). Such a discriminant position is called *salient* and is specific for each object o_n. Section 8.1 proposes the calculation of networks of salient regions for the selection of the next camera position in 2D. In this context the active recognition algorithm can be seen as fixation control for recognition. Section 8.2 proposes the selection of salient viewpoints based on information theory in order to define an active object recognition algorithm as viewpoint planning. Experimental results are given for the Columbia database with 100 objects seen from 72 viewpoints. The most interesting result is that the recognition results are 3D–consistent. This is remarkable since the algorithm is entirely based on the appearance of objects without explicitly incorporating 3D information. However, the experimental results of sections 8.1 and 8.2 are preliminary and should be seen as such.

An important question for fixation control in an active vision system is characterized by the question *"where to look next?"* [Swa 93b]. We can divide the different approaches to this problem into three categories: task driven, context driven and feature driven attention mechanisms. In this chapter we propose a new statistical approach to answer the question *"where to look next?"*. Our approach is based on the use of multidimensional histograms of local neighborhood operators and a network of salient points. The multidimensional receptive field histograms serve as a basis for the generation of hypothesis for the presence of an object. The network of salient points (respectively salient viewpoints) of objects enables the technique to move the camera between regions of interest (respectively to points of view). The technique therefore drives fixation in order to verify the generated hypothesis. Since we have an expectation about what we should observe at the new camera position the approach allows us to also answer the question of *"what to look for?"*.

Based on the preceding chapter 7, section 8.1 develops a technique that obtains the network of salient points and uses this network for fixation control in the context of active object recognition. Some examples of the calculation of salient points and their robustness to scale changes are given. Section 8.2 proposes active object recognition as viewpoint planning. Salient viewpoints are selected through maximization of the transinformation over different viewpoints of an object. Experimental results are given for the Columbia image database of a 100 objects.

8.1 Active object recognition in a single image

In the previous chapter we extended the use of multidimensional receptive field histograms for probabilistic object recognition. In this section we reverse the object recognition problem by asking the question, *"where should we look?"*, when we want to verify the presence for object, to track an object or to actively explore a scene. This section describes a method for deriving a network of salient points for an object which are characteristic for an object. These salient points are literally the points on the object which are the almost unique. These points maximize the discrimination between objects. A network of salient points may be used to control fixation in the context of object recognition.

8.1.1 Network of salient points

This first part of the section introduces a concept to determine the most significant points in an image and/or of an object. We use this concept to determine a network of salient points for a set of objects and use these networks in the context of fixation control for object recognition (see section 8.1.2). It is worth mentioning that this concept generalizes to an *interest point detector*. Such interest points for an object or in an image can be for example used directly in a geometric hashing approach for object recognition [Wol 90] (see section 2.2.1). We may also use the *interest point detector* to find track-able (that is discriminable and easy to detect) points of an object.

As described in section 7.1 we can calculate the probability of an object o_n, given a local measurement vector m_k:

$$p(o_n|m_k) \quad = \quad \frac{p(m_k|o_n)p(o_n)}{p(m_k)} \tag{8.109}$$

with

- $p(o_n)$ the a priori probability of the object o_n,
- $p(m_k)$ the a priori probability of the filter output combination m_k,
- $p(m_k|o_n)$ the probability density function of object o_n, which differs from the histogram of local measurement vectors for an object o_n only by a normalization factor.

In this section we are interested in the most significant points of a certain object o_n (or a given image). These points can be obtained by maximizing $p(o_n|m_k)$ over all filter outputs m_k of the object o_n. Therefore we maximize the relation between $p(m_k|o_n)$ and $p(m_k)$ (as we will see later, we may neglect $p(o_n)$). This results in highly discriminant and, as the examples show, easy to detect filter combinations.

Such salient points (or interest points) can be used in many ways: since these points are the most discriminant we can use them in the context of active vision (as for example for tracking or fixation control (section 8.1.2)). In the context of object recognition we can use the first K maxima in order to minimize the number of points needed to determine the identity of an object. Such K maxima define a network of salient points, which we will use in the following section in the context of an active recognition system.

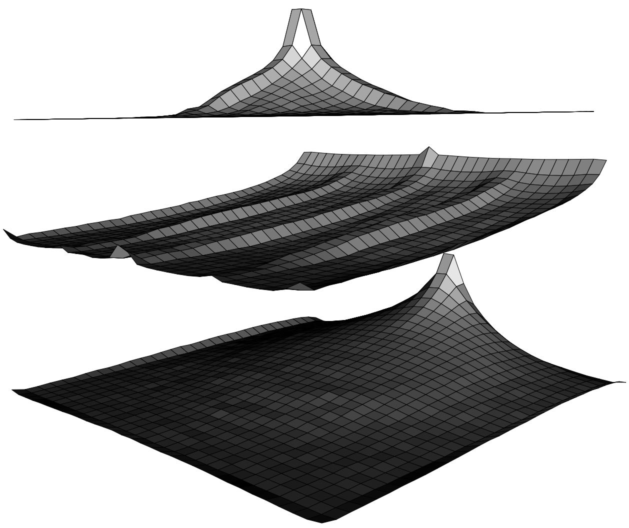


Figure8.1: Average Histograms for different filter combinations: top Dx-Dy, middle Mag-Dir and bottom Mag-Lap ($\sigma = 1.0$, number of bins per histogram axis is 32).

In order to use equation 8.109, the a priori probability $p(m_k)$ of each local filter output m_k has to be determined. This probability distribution depends not only on the used filter combination but also on the context of the vision task. Therefore the calculation of the probability distribution $p(m_k)$ needs careful investigation.

In chapter 7 we used the equation $p(m_k) = \sum_i p(m_k|o_i)p(o_i)$ to calculate the a priori probability of the measurement vector m_k. This equation works well in the context of probabilistic object recognition, but depends on the given database and considered backgrounds. Such a dependency can be justified if the environment of the visual task is known in advance. This is often the case in a tracking task, where the camera is stationary. Here we want to deal with the more general case, where not all objects/backgrounds are known in advance. Therefore the *real* a priori probability $p(m_k)$ has to be approximated or estimated. We propose the calculation of the average histogram of a sufficiently large representative image database. Figure 8.1 shows three such average 2D-histograms for the filter combinations *Dx-Dy*, *Mag-Dir* and *Mag-Lap*, which have been calculated on a database of 832 images. Even though this average histogram still depends on the database which is used, it seems to be more reliable than a purely analytical approximation.

Using an average histogram $p(m_k|average)$ for the approximation of the a priori probability $p(m_k)$ we can rewrite equation 8.109 (in which the probability $p(o_n|m_k)$ depends only on the considered object o_n and the average histogram):

$$p(o_n|m_k) = \frac{p(m_k|o_n)p(o_n)}{p(m_k|average)} \tag{8.110}$$

We may omit the probability $p(o_n)$ since its value does not affect the order of the maxima of $p(o_n|m_k)$. In order to stabilize the equation to outliers we add the histogram $h(m_k|o_n)$ on top of the average histogram $h(m_k|average)$. We use therefore the following equation to determine the most salient points of the object o_n:

$$p^\star(o_n|m_k) = \frac{h(m_k|o_n)}{h(m_k|o_n) + h(m_k|average)} \tag{8.111}$$

In this equation $h(m_k|o_n)$ is the histogram value of the measurement vector m_k for the object o_n and $h(m_k|average)$ is the histogram value of the measurement vector m_k in the average histogram.

By using the first K maxima of equation 8.111, we obtain a network of salient points as well as a quantitative measure for the significance of each point (in particular $p^\star(o_n|m_k)$). This measure can be used to determine if a certain object contains points or regions which are sufficiently discriminant. The higher the value $p^\star(o_n|m_k)$ of the salient points the more they are discriminant. In section 8.1.3 we will use these values to determine objects in a database which contain the most discriminant and objects with the least discriminant points.

Figure8.2: An object and its first 8, 50 and 103 salient points

Figure 8.2 shows an example of the first salient points of an object (marked with circles). Table 8.1 shows the corresponding filter output values. Here we used the 3D histogram *Dx-Dy-Lap* with $\sigma = 1.0$. The size of the circles has been chosen arbitrarily and does not correspond to the support of the Gaussian filters. The maxima in table 8.1 correspond to high filter output values, which makes them not only discriminant but also easy to detect in an image. They correspond to a very low value of $h(m_k|average)$ (the first maxima for example occurs in average only 0.04 times per image). This results in a very high value for $p^\star(o_n|m_k)$.

In many cases we obtain cluster of salient points (as for example in figure 8.2) so that it seems convenient to define a network of salient regions rather than a network of salient points. The concept for fixation control described below can be easily extended to such a network of salient regions.

Number of Maxima	Dx	Dy	Lap	$h(m_k \mid o_n)$	$h(m_k \mid average)$	$p^\star(o_n \mid m_k)$
1	34.2	49.0	-99.0	1	0.0445	0.9574
2	-19.4	-50.5	56.8	4	0.2067	0.9509
3	-27.3	-49.0	63.7			
4	-26.6	-48.7	54.4			
5	-25.5	-53.1	62.7			
6	36.0	50.8	6.2	3	0.1635	0.9483
7	35.8	50.7	7.8			
8	40.1	48.2	3.3			
...						
103	16.1	43.8	16.0	19	1.2848	0.9367

Table8.1: The filter output values (scaled into the interval $[-128, +128]$) for the first salient points of the object in figure 8.2, the values of the histogram of the object $h(m_k|o_n)$, the average histogram $h(m_k|average)$ and the calculated value $p^\star(o_n|m_k)$.

8.1.2 Fixation control for object recognition

The general scheme which we propose for active fixation control in the context of object recognition can be divided in two major steps:

Hypothesis generation: Based on the probabilistic object recognition algorithm introduced above we generate a hypothesis for the presence of an object at a certain scale and rotation. After a local scan for a salient point of the best hypothesis objects, we also obtain a hypothesis for the objects position.

Verification: In order to verify the generated hypothesis, we can use the *network of salient points* of this object to determine the next region of interest (*where to look next*). Having this network of salient points we also know *what to look for*.

Hypothesis generation is mainly based on the described probabilistic object recognition algorithm. An interesting property of the algorithm is the ability to calculate probabilities for each object at an arbitrary image location. This means that we can chose the initial region of interest arbitrarily and obtain a hypothesis concerning the presence of an object. Since we assume that the objects are in the database at different scales and rotations we obtain not only the most probable objects but also their most likely scale and rotation.

By choosing the most probable objects we can scan locally the actual region of interest in order to find a salient point for an object. By doing so we incorporate calculated local characteristics to gather evidence for the chosen objects. Once a salient point for one of the most probable objects is found we hypothesize this object. The position of the salient point in the corresponding network of salient points is used to determine the approximate position of the object in the image.

For *verification* of the generated hypothesis (which contains information about the most probable object with its scale, rotation and position in the image) we want to choose the next region of interest in the image, which is most suited for the verification step. Using the network of salient points of the hypothesized object we can choose the most significant (or the next most significant) salient point in the network to determine the next region of interest (*where to look next*). Since we know the vector of local characteristics for this salient point we also know *what to look for* in the new region of interest.

Finding the expected salient point, we can incorporate one or several verification steps until the evidence for the object is sufficient to *confirm* the hypothesis. If we cannot find the expected salient point in the new region of interest, we can incorporate the observed local characteristics to calculate and/or modify the probabilities of each object (*recover*).

In the context of the thesis we did not implement the full algorithm due to time constraints. Section 8.1.3 describes illustrative examples of the application of the salient point detector and gives some indication of its robustness to scale changes. The next step for the realization of the defined fixation control algorithm would be an analysis of the repeatability and stability of the salient point detector with respect to scale changes, image plane rotation and light changes. Additionally, a representation of the network of salient points to be defined and an efficient algorithm has to be used for the local search of salient points. In particular, the repeatability as well as the stability of the salient point detector could limit the applicability of the proposed approach.

8.1.3 Examples for illustration of the salient point detector

In this section we want to apply the detector of salient points to a database of 30 objects shown in figure A.1. Throughout the section we use six–dimensional histograms, namely the filter combination *Dx-Dy-Lap* at two different scales ($\sigma = 1.0$ and 2.0). By using the first 30 maxima of $p^\star(o_n|m_k)$ (see equation 8.111) we obtain a network of salient points for each object. In the first part of the section we determine the objects with the most discriminant and with the least discriminant salient points. In the second part we examine the robustness of the salient point operator to scale changes.

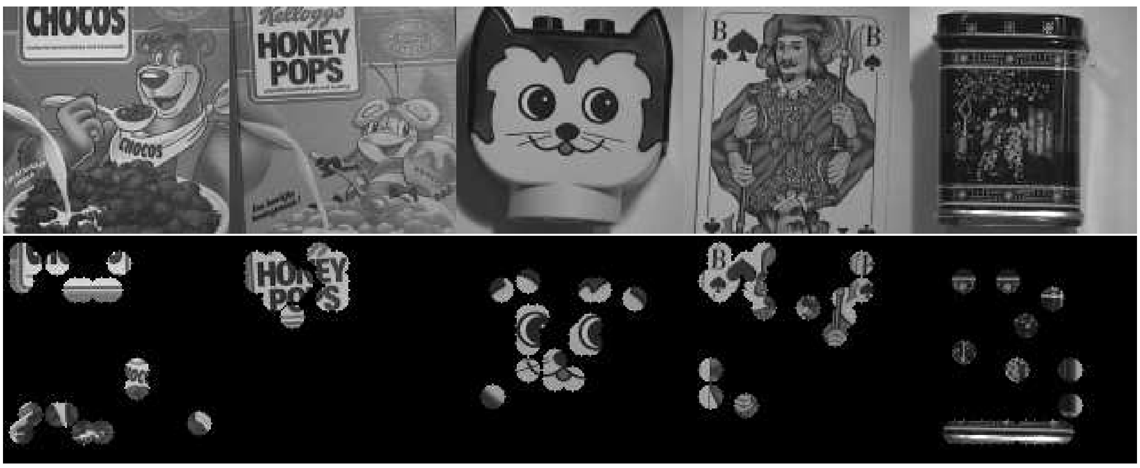


Figure8.3: The first row shows the objects which contain the most discriminant points. The second row shows the 20 most salient points for these objects

As mentioned in section 8.1.1 we can use the value of $p^\star(o_n|m_k)$ to determine the objects of the database which contain the most discriminant points. The first row of figure 8.3 shows the

Figure8.4: The 5 objects which contain the least discriminant points. The second row shows 20 most salient points of these objects. This represents a worst case for application of the technique.

5 objects which contain the most discriminant points. The second row of this figure shows the 20 most discriminant points of each object. In figure 8.4 we show the 5 objects of the database which contain the least discriminant points, corresponding to the worst case for application of the technique. The second row shows the 20 most salient points for these objects.

Figure8.5: Robustness of the salient point detector to scale. The scale between each image is approximately 20%

In order to examine the robustness of the salient point detector with respect to scale changes we used one of the objects with the most discriminant points (figure 8.3) and one of the objects with the least discriminant points (figure 8.4). Figure 8.5 shows the result of the same detector of salient points for three different scales (of each object). The approximate scale change between each image is 20%. We can see that the salient points are almost the same, indicating a certain robustness of the salient point detector to scale changes.

8.2 Active object recognition as viewpoint planning

The previous section defined an active object recognition algorithm in the context of fixation control of the camera. Camera movements are therefore in 2D, that is to say within the image plane. Here we propose the calculation of salient viewpoints based on the transinformation of viewpoints. This enables the definition of an active object recognition algorithm in the context of viewpoint planning.

8.2.1 Transinformation for a single viewpoint of an object

In section 4.2 we showed the application of information theoretical concepts to the process of object recognition. In this section we want to go a step further in applying the developed concepts in the context of viewpoint planning for object recognition. More precisely we will calculate the transinformation of each viewpoint of an object in order to choose the most *salient viewpoints* of an object. Section 8.2.2 will use these viewpoints in an active object recognition approach.

In section 4.2.3 we defined the transinformation of the object recognition process based on the event pairs (o_n, m_k). In the following we will develop a equation for the choice of the most salient viewpoint of an object. Starting from the original equation 4.73 we can rewrite the transinformation:

$$\begin{aligned} T(O,M) &= \sum_{n=1}^{N}\sum_{k=1}^{K} p(o_n \wedge m_k) \log \frac{p(o_n \wedge m_k)}{p(o_n)p(m_k)} && (8.112)\\ &= \sum_{n=1}^{N} p(o_n) \sum_{k=1}^{K} p(m_k|o_n) \log \frac{p(m_k|o_n)p(o_n)}{p(o_n)p(m_k)} && (8.113)\\ &= \sum_{n=1}^{N} p(o_n) \sum_{k=1}^{K} p(m_k|o_n) \log \frac{p(m_k|o_n)}{\sum_i p(m_k|o_i)p(o_i)} && (8.114) \end{aligned}$$

The transinformation can be interpreted as the average transinformation for an object o_n which may be defined as:

$$T(o_n, M) = \sum_{k=1}^{K} p(m_k|o_n) \log \frac{p(m_k|o_n)}{\sum_i p(m_k|o_i)p(o_i)} \qquad (8.115)$$

The probability $p(m_k|o_n)$ corresponds to the probability density function of an entire object as introduced previously. Since we want to calculate the transinformation of each individual viewpoint of an object we define the transinformation of an object o_n at a certain pose $S_j = (t_z r_x r_y r_z)$ as:

$$T(o_n, S_j, M) = \sum_{k=1}^{K} p(m_k|o_n, S_j) \log \frac{p(m_k|o_n, S_j)}{\sum_{i,l} p(m_k|o_i, S_l)p(o_i, S_l)} \qquad (8.116)$$

8.2.2 Viewpoint planning

This section describes an active object recognition algorithm. The principal idea of the algorithm is to hypothesize the object identity and its pose from a test-image (single view recognition). Based on this estimate the algorithm moves the camera to the most discriminant viewpoint(s) of the hypothesized object. The information gathered from the new viewing direction is then used for the verification of the hypothesized object and its pose.

Experimental results (section 8.2.3) show that the algorithm can resolve all (except one) of the ambiguities inherent to a single–view recognition algorithm. Most interestingly the algorithm incorporates 3D information from the statistical representation of the object. Remaining errors (for one verification step) are therefore 3D consistent. By applying multiple verification steps the algorithm incorporates more 3D information which results in a nearly 100% recognition rate.

Outline of the active object recognition algorithm

As motivated in the introduction of this chapter, the outline of the active recognition algorithm contains three steps – here we use histogram matching for the generation of hypotheses:

Hypothesis generation: the hypothesis is generated by matching the histogram $H(M|I)$ of the test images I with the database of histograms. An object and its pose is hypothesized, if the corresponding multidimensional receptive field histogram produces the minimum of the χ^2_{qv}-statistics: the hypothesized object is $\hat{o_n}$ with its pose parameters $\hat{S_j}$ (see equation 8.108).

Camera movement: the camera is moved to the most (or second most) discriminant viewpoint of the object $\hat{o_n}$. This discriminant viewpoint is calculated off-line on the basis of equation 8.116. The movement of the camera is calculated on the basis of the difference ΔS between the current estimated pose $\hat{S_j}$ and the pose of the most discriminant viewpoint.

Verification of the hypothesis: at the new camera position we will have (see equation 8.108) again a hypothesized object and its hypothesized pose. Different conditions can be used for the verification step. First of all the hypothesized object should be the same before (time $(t-1)$) and after (time (t)) the camera movement:

$$\hat{o_n}(t) = \hat{o_n}(t-1) \tag{8.117}$$

Furthermore we can use knowledge of the camera movement which has been calculated on the basis of the difference ΔS. By using these differences we can predict the observation at the new camera position and compare them to the obtained hypothesis:

$$\hat{S_j}(t) \quad = \quad \hat{S_j}(t-1) + \Delta S(t-1) \pm \epsilon \tag{8.118}$$

ϵ corresponds to the allowed error in pose estimation.

The proposed active object recognition algorithm moves the camera to the most discriminant viewpoint of the hypothesized object. By doing so we can expect to verify the hypothesized object if it is the correct object. On the other hand, we expect to be able to reject the hypothesized object whenever it is not the correct one. We can expect this as the most discriminant viewpoints are chosen relative to the database. Ambiguities in the database, which cannot be solved by a single–view recognition system are expected to be solvable by the proposed approach.

The most attractive property of the proposed active object recognition process is that not only 2D information (a single image) but also 3D information of the probability density function of the objects is used in order to verify the hypothesized object. That implies that errors of the proposed approach should only be possible between objects which are 3D–compatible (for example different cubes). In the experiments described below we observe this property for the errors in the system.

Being interested in an object recognition process with minimal false positive and minimal false negative identifications at the same time, we can introduce several verification steps. If we use L verification steps we obtain the following conditions for the verification step:

$$\forall t = 2, ..., L : \hat{o_n}(t) = \hat{o_n}(t-1) \quad (8.119)$$

$$\hat{S}_j(t) = \hat{S}_j(t-1) + \Delta S(t-1) \pm \epsilon \quad (8.120)$$

8.2.3 Experimental results

This section describes an experiment designed to show the applicability of the proposed active object recognition approach. The experiments are based on the Columbia image database of 100 objects [Nen 96], which contains 7200 image, 72 for each object (originally the database contains color images these were converted to grey scale images). The images are taken under controlled lighting conditions in front of a black background. Figure A.6 shows the 100 objects. The 72 viewpoints of an object are taken from a fixed camera position where the object has been turned in 5° intervals on a turntable.

Unfortunately the database contains only one rotational degree of freedom. Consequently three parameters of the pose estimation are not considered in the experiment. The reason for using this database was the possibility of simulating camera movement by "turning" the object in front of the camera. Therefore we can validate the proposed algorithm without the depending on camera movement and camera calibration (in order to move the camera relative to the object we have to calibrate the camera at least approximately).

Half of the images (every 10°) of the objects ($100 \times 36 = 3600$ images) are used as a database. The remaining half of the images are taken as a test set. For each database image we calculate its histogram of the first Gaussian derivative in x– and y–directions. In this particular experiment we used a resolution of 16 cells per histogram axis so that each of the 3600 images is represented by $16^2 = 256$ numbers. For each object we have calculated the most and the second most discriminant viewpoint.

verification steps	% recognition	number of errors
0	98.36	59
1	99.14	31
2	99.97	1

Table8.2: Experimental results on the Columbia images database of 100 3D-objects. For comments see text.

Table 8.2 shows the results which we obtained by the application of the active object recognition algorithm introduced above. The first row shows the number of verification steps which we used in order to accept a hypothesized object. Without using any verification (first line of the table) we obtain 59 misclassifications of the 3600 test images. By using only one verification step we can reduce this number to 31 misclassifications. By using 2 verification steps we can resolve nearly all ambiguities and we obtain only one misclassification. These results validate the applicability of the proposed active recognition approach. (During the verification steps we allowed an error ϵ of up to $5°$ for the involved rotational pose parameter, see equation 8.120).

Figure8.6: The 5 objects which are confused by the active recognition algorithm with one verifications step

It is interesting to note that all misclassifications produced with one verification step are 3D consistent. Figure 8.6 shows the five objects which have been confounded. Most errors (20 of 31) are made between the last two objects (which can be distinguished in the original database by their color). All five objects are 3D consistent because they are all cuboids. Therefore the errors are not arbitrary but rather "systematic". We may discriminate these objects by using color information or other appropriate measurement vectors (which can be evaluated by the transinformation introduced in section 4.2.3). The only misclassification obtained with two verification steps is between the last two object of figure 8.6.

8.3 Conclusion

This chapter proposed the application of multidimensional receptive field histograms for active object recognition. Two algorithms have been proposed namely active object recognition as fixation control and active object recognition as viewpoint planning.

In the case of recognition as fixation control, we have shown how multidimensional receptive field histograms can be used to deduce the measurement vectors which are maximally discriminant (= salient) for each object within a set. Networks of such discriminant points may be used to determine *"where to look next?"* and *"what to look for"* in the context of active object recognition. In particular a salient point detector has been defined which can be used in a general way for example for geometric hashing.

Salient viewpoints can be chosen by maximizing the transinformation over viewpoints of an object. The salient viewpoint selection is based on an analogy developed in section 4.2 between information theory and object recognition. Salient viewpoints can be used in the context of active object recognition by moving the camera to the most discriminant viewpoints of a hypothesized object. Even though preliminary, the experimental results on the Columbia image of 100 objects indicate that 3D–consistent recognition can be obtained based on the proposed algorithm. This is remarkable since the underlying appearance based algorithm does not explicitly introduce 3D measurements.

In particular, the promising results of this and the last two chapters (chapter 6: reliable object recognition by histogram matching, chapter 7: robust probabilistic object recognition and here 3D–consistent object recognition) allow us to propose an *object classification* framework based on multidimensional receptive field histograms in the following chapter 9, introducing the concept of visual classes.

Chapter 9

Object classification

The preceding chapters have shown that multidimensional receptive field histograms provide a reliable means for the recognition of 3D objects. Three different algorithms have been proposed for the recognition of objects: histogram matching, probabilistic object recognition and in the previous chapter an active object recognition algorithm. Recognition results from these chapters indicate that visually similar objects have similar multidimensional receptive field histograms. This leads to the conclusion that similar multidimensional receptive field histograms may correspond to visually similar objects. Most noteworthy here is the 3D–consistent recognition achieved by incorporating multiple views of the same object as proposed for active object recognition. These results permit us to introduce the concept of visual classes as a general framework for object classification based on appearance. Classification is one of the principal challenges for object recognition techniques based solely on 2D–image measurements.

The first section of the chapter introduces the concept of *visual classes* as a general framework for object classification. Visual classes associate appearances which are similar with respect to a set of image measurements. As defined here, visual classes are implicit in many object representation schemes (geometric as well as appearance based models). We argue that the identification of visual classes provides a powerful tool for object classification. They provide a first step to object classification which depends on information provided to the observer, including context dependency and relations in space and time between objects.

Based on the statistical representation of objects by multidimensional receptive field histograms, section 9.2 proposes a maximum–likelihood technique for the recognition of visual classes. Section 9.3 describes experiments of the retrieval of visually similar images from a database of 200 countryside images. The important question of how to extract visual classes from a histogram database is discussed in section 9.4. In particular, the application of a special clustering algorithm is proposed.

9.1 The concept of visual classes

The section motivates the concept of *visual classes* for object classification. Visual classes are defined on the basis of 2D– and/or 3D–similarities in appearance. Therefore, visual classes depend only on image content. We argue that the identification of visual classes provides a powerful tool for object classification. Visual classes, defined with information which is intrinsic to the image, provide a first step to object classification which depends on information provided to the observer, including context dependency and relations in space and time between objects. We anticipate that visual classes will be suitable for generalization beyond a given database. So that we can identify the visual class (and therefore the possible object classes) of an unknown object.

In previous chapters we have developed and applied a statistical object representation for arbitrary 3D objects based on multidimensional receptive field histograms. This statistical object representation is compatible with many object representation schemes. The representation may be used to identify 2D– and/or 3D–similarities of 3D objects. As we have seen in the previous chapter multidimensional receptive field histograms may provide a means for 3D–consistent object recognition. Based on visual similarities between objects we define the concept of visual classes.

Visual classes associate appearances which are similar with respect to a set of image measurements. Visual classes should be based entirely on the image or image measurements. The implication is that visual classes typically *do not* coincide with object classes as they would be defined in the context of a certain task. A well known example is the object class of CHAIRS, which happens to posses many different appearances. The typically approach is to find a single representation for all (or at least most) appearances of the object class CHAIR (often called invariant representation). We argue that such a representation does not exist in general but that we can define visual classes for different appearances of the object class CHAIR. An object class therefore defines several visual classes. On the other hand it is likely that a single visual class can be shared between different object classes.

We argue that visual classes (as defined here) can be seen as a subproblem of object classification. Starting from visual classes, we can define (or learn) object classes as a collection of visual classes. In general an object class will be defined by a combination of different visual classes. Possible combinations describe relations between objects and depend upon the context (time and space), and other knowledge imposed by an observer. Since it has been proven very difficult to extract object classes directly from the image we propose to extract visual classes from an image as a preliminary step for object classification. The concept of visual classes therefore should be seen as an attempt to fill the gap between appearance based object identification and a general object classifier.

The statistical representation introduced in section 4.1 is seen as a general framework for the representation of objects since it is compatible with many different object representations. As mentioned above, visual classes are defined by similarities (2D and 3D) with respect to a set of image measurements. Such similarities (in 2D and/or 3D) may be extracted from a statistical object representation by dimensionality reduction schemes such as the Karhunen–Loeve transformation, discriminant analysis and clustering algorithms. In section 9.2 we will introduce a possible approach to representation and recognition of visual classes in a maximum–likelihood framework. Section 9.3 describes first experiments validating the proposed concept of visual classes.

9.2 Recognition of visual classes

In the following we describe the maximum likelihood recognition of visual classes. The idea is to adapt the algorithm of Moghaddam and Pentland [Mog 95] who propose a maximum likelihood framework for the visual search and detection of faces and hands. Their approach is based on the estimation of the probability density for a vector $\mathbf{x}$ given a class Ω: $p(\mathbf{x}|\Omega)$. In their case the vector $\mathbf{x}$ is a gray-level image and the class Ω is either a face, a part of a face (eye, mouth or nose) or the silhouette of a hand.

We can adapt their approach for the recognition of a *visual class* Ω by using a vector representation for multidimensional receptive field histograms as vector $\mathbf{x}$. In the following we assume that we have extracted K histograms $H_1, H_2, \ldots, H_K$ from a database of multidimensional receptive field histograms representing a visual class Ω. Each histogram corresponds to a particular appearance of a 3D object. The vector representation of a histogram H_i is given by the transformation t: $\mathbf{x}_i = t(H_i)$. Each cell of the histogram H_i therefore corresponds to one dimension of the vector $\mathbf{x}_i$.

In order to estimate the probability density function $p(\mathbf{x}|\Omega)$ we assume that we have robustly estimated the mean vector $\overline{\mathbf{x}}$ and the covariance matrix $\Phi = \sum_{i=1}^{K}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$ from the distribution of the samples $\mathbf{x}_i = t(H_i)$. If we further assume that $p(\mathbf{x}|\Omega)$ is a Gaussian distribution we obtain:

$$p(\mathbf{x}|\Omega) = \frac{e^{-\frac{d(\mathbf{x})}{2}}}{\sqrt{(2\pi)^N ||\Phi||}} \tag{9.121}$$

with $d(\mathbf{x})$ the Mahalanobis distance:

$$d(\mathbf{x}) = (\mathbf{x} - \overline{\mathbf{x}})^T \Phi^{-1} (\mathbf{x} - \overline{\mathbf{x}}) \tag{9.122}$$

In order to rewrite the Mahalanobis distance we calculate the eigenspace of the samples $\mathbf{x}_i$. By projecting all K histograms $H_1, H_2, \ldots, H_K$ onto the vectors $\mathbf{x}_i = t(H_i), i = 1, 2, \ldots, K$ and by using the mean vector $\overline{\mathbf{x}}$, the covariance matrix is given by:

$$\Phi = \sum_{i=1}^{K} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \tag{9.123}$$

This covariance matrix provides a series of eigenvalues λ_i and eigenvectors e_i (see for details of computation [Pre 92, Tur 91a]):

$$[e_1 e_2 \dots e_N] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_N \end{bmatrix} = E\Lambda = \Phi E \tag{9.124}$$

Using all N eigenvectors we can define a orthonormal basis $E = [e_1 e_2 \dots e_N]$ and can project each histogram H_i onto this eigenspace. The projection $\mathbf{y}_i$ is given by:

$$\mathbf{y}_i = E^T(\mathbf{x}_i - \overline{\mathbf{x}}) \tag{9.125}$$

The eigenvector basis E and the eigenvalue matrix Λ allow us to rewrite the Mahalanobis distance:

$$\begin{aligned} d(\mathbf{x}) &= (\mathbf{x} - \overline{\mathbf{x}})^T \Phi^{-1} (\mathbf{x} - \overline{\mathbf{x}}) & (9.126) \\ &= (\mathbf{x} - \overline{\mathbf{x}})^T E \Lambda^{-1} E^T (\mathbf{x} - \overline{\mathbf{x}}) & (9.127) \\ &= \mathbf{y}^T \Lambda^{-1} \mathbf{y} & (9.128) \\ &= \sum_{i=1}^{N} \frac{y_i^2}{\lambda_i} & (9.129) \end{aligned}$$

with $\mathbf{y} = E^T(\mathbf{x} - \overline{\mathbf{x}})$ the projection of the vector $\mathbf{x}$ into the eigenspace. By calculating only the first M eigenvector projections Moghaddam and Pentland propose the following estimate of the Mahalanobis distance $d(\mathbf{x})$:

$$\begin{aligned} \hat{d}(\mathbf{x}) &= \sum_{i=1}^{M} \frac{y_i^2}{\lambda_i} + \frac{1}{\rho} \sum_{i=M+1}^{N} y_i^2 & (9.130) \\ &= \sum_{i=1}^{M} \frac{y_i^2}{\lambda_i} + \frac{1}{\rho} \quad \epsilon^2(\mathbf{x}) & (9.131) \end{aligned}$$

The space, spanned by the first M eigenvectors is called *feature space* F in the following. The complement of this space, $\overline{F}$ is spanned by the remaining $N - M$ eigenvectors. Based on the estimated distance $\hat{d}(\mathbf{x})$, the probability $p(\mathbf{x}|\Omega)$ can be estimated as the product of two marginal independent Gaussian densities, namely by $p_F(\mathbf{x}|\Omega)$ the probability density function within the *feature space* F and the estimated probability density function $\hat{p}_{\overline{F}}(\mathbf{x}|\Omega)$ for the complement space $\overline{F}$:

$$\begin{aligned} \hat{p}(\mathbf{x}|\Omega) &= p_F(\mathbf{x}|\Omega) \quad \times \quad \hat{p}_{\overline{F}}(\mathbf{x}|\Omega) & (9.132) \\ &= \frac{e^{-\frac{1}{2}\sum_{i=1}^{M} \frac{y_i^2}{\lambda_i}}}{\sqrt{(2\pi)^M \prod_{i=1}^{M} \lambda}} \times \frac{e^{-\frac{\epsilon^2(\mathbf{x})}{2\rho}}}{\sqrt{(2\pi)^{(N-M)} \rho^{(N-M)}}} \end{aligned}$$

The first part of the estimated distance is called the *distance in feature space* F since it lies on the hyper-plane F spanned by the M eigenvectors. The distance $\epsilon^2(\mathbf{x})$ is called *distance from feature space* and can be calculated directly from the eigenvector projection:

$$\epsilon^2(\mathbf{x}) = \sum_{i=M+1}^{N} y_i^2 = ||(\mathbf{x} - \overline{\mathbf{x}})||^2 - \sum_{i=1}^{M} y_i^2 \tag{9.133}$$

An optimal solution for the parameter ρ with respect to the relative entropy – as cost function – between the true density $p(\mathbf{x}|\Omega)$ and the estimated density $\hat{p}(\mathbf{x}|\Omega)$ is given by:

$$\rho^\star = \frac{1}{N-M} \sum_{i=M+1}^{N} \lambda_i \tag{9.134}$$

Moghaddam and Pentland extend the approach to multimodal F–space densities. Under the assumption that the $\overline{F}$–space components are independent of the components of the F–space we can still use the separable form of the density estimate $\hat{p}(\mathbf{x}|\Omega)$ of equation 9.132. $p_F(\mathbf{x}|\Omega)$ then becomes an arbitrary density in the principal components of $\mathbf{y}$. A possibility is to use a mixture of multimodal Gaussian densities, estimated by the Expectation maximization algorithm (see also section 2.3.2).

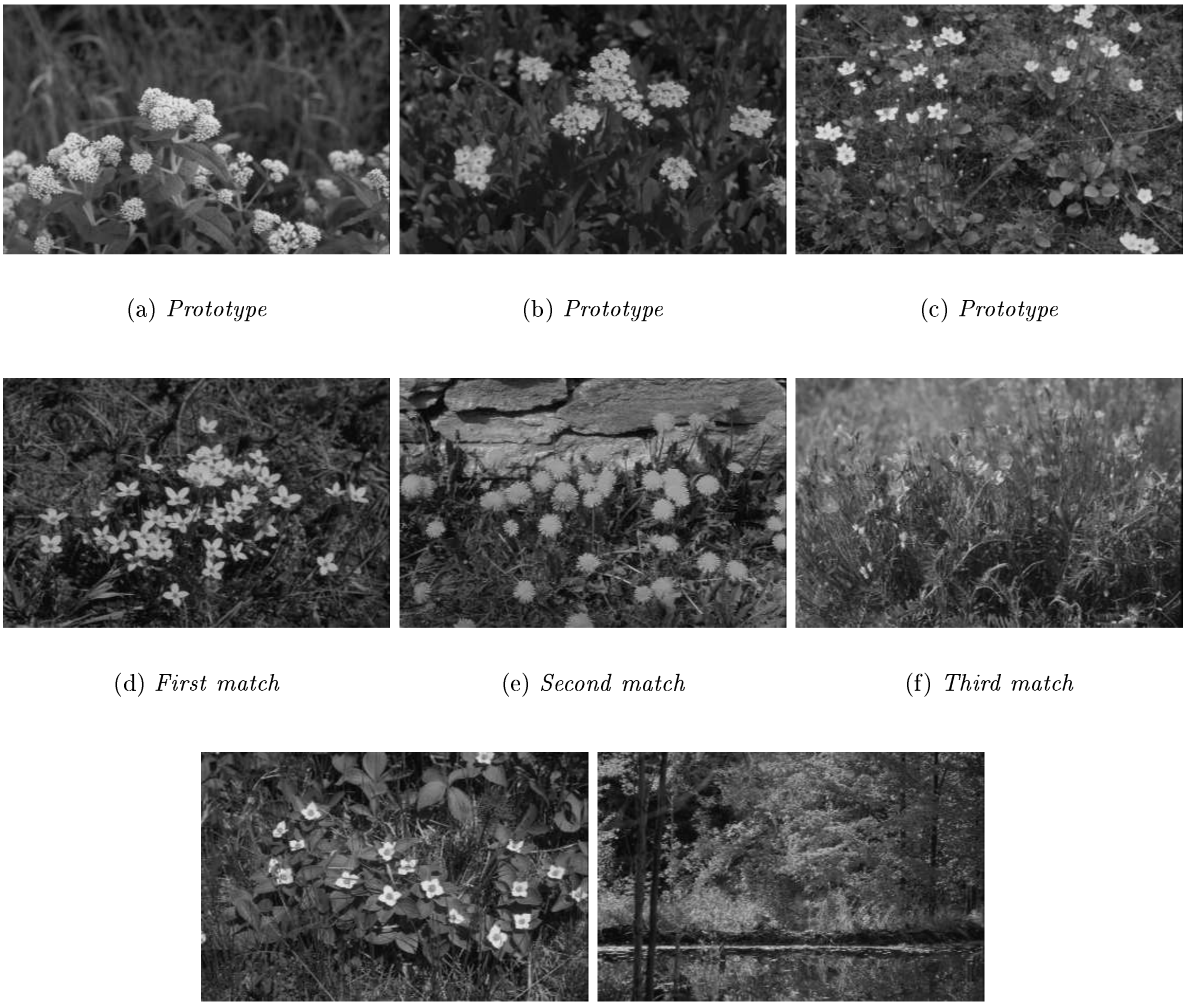
The proposed combination of distance from feature space and distance in feature space provides a statistical method for the maximum likelihood recognition of visual classes. In the following section we apply in particular the *distance from feature space* $\epsilon^2(\mathbf{x})$ for the classification of images.

9.3 Classification examples

The previous section introduced the *distance from feature space* and the *distance in feature space* for the recognition of visual classes. This section shows examples of the recognition of the visual class FLOWER and the visual class COAST. In particular, we choose three images per visual class and calculate the two–dimensional eigenspace spanned by the three histograms of the images. For finding visually similar images we compute the distance from feature space of 200 images[1] with respect to the eigenspace. The database contains images of mountains, flower, countrysides, forests, trees and the like. For each of the 200 images we compute a histogram. The *distance from feature space* $\epsilon^2(\mathbf{x})$ of these histograms is used in order to decide if an image is visually similar to a visual class.

The filter combination used is given by *Dx-Dy-16* (first Gaussian derivatives in x– and y–directions at a resolution of 16 cells per histogram axis) at three different scales, namely $\sigma_1 = 1.0, \sigma_2 = 2.0$ and $\sigma_3 = 4.0$. The corresponding multidimensional receptive field histograms are therefore six-dimensional. For each of the 200 images in the image database we calculate one single six-dimensional histogram. Three histograms are used for the computation of the two–dimensional eigenspace of a visual class as for example in the case of the visual class FLOWER.

[1] The image database used for the experiments has been provided by Jutta Kreyß of the University of Bremen, Germany



(a) *Prototype* (b) *Prototype* (c) *Prototype*

(d) *First match* (e) *Second match* (f) *Third match*

(g) *Fourth match* (h) *Fifth match*

Figure9.1: Flower recognition. The first three images (a), (b) and (c) have been used to define the visual class FLOWER *and the remaining images correspond to the five best matches of a database of 197 images*

Figures 9.1(a), (b) and (c) show the three manually chosen images of the image database for the representation of the visual class FLOWER. By using the distance from feature space $\epsilon^2(\mathbf{x})$ we can order the remaining 197 images. Figures 9.1(d) through (h) show the first five matches obtained. The four best matches are all flowers, which shows the applicability of distance from feature space for the recognition of the visual class FLOWER. The fifth match is visually similar because of the grass structure contained in all three prototypes of the class. We have to point out that not all combinations of three flower images result in the same quality of results that is to say retrieving flower images in most cases. An appropriate algorithm has to be defined for the extraction of images which are suitable for the definition of a visual class.

(a) *Prototype* (b) *Prototype* (c) *Prototype*

(d) *First match* (e) *Second match* (f) *Third match*

(g) *Fourth match* (h) *Fifth match*

Figure9.2: Coast recognition. The first three images (a), (b) and (c) have been used to define the visual class COAST *and the remaining images correspond to the five best matches of a database of 197 images*

Figure 9.2 shows a second example for the recognition of a visual class, in particular, the recognition of the class COAST. Once again three coastal images (see figure 9.2 (a), (b) and (c)) have been chosen manually and used for the computation of a two–dimensional eigenspace of the class COAST. The five closets matches are visually similar and all show coastal scenes.

The appendix shows three further examples of the application of distance from feature space for the recognition of visual classes. The examples use the same histograms and the same image database of 200 images. In particular, the recognition of the visual class forest images (see figure B.3), glacier images (see figure B.4) and waterfall images (see figure B.5) are shown. In all three cases the five closest matches correspond to images containing visually similar scenes. In the case of the forest and the glacier recognition all five matches are forest and glacier scenes respectively. In the case of waterfall recognition, all three waterfall images (of the remaining 197 images) are contained in the first four matches.

The results obtained here in the context of image retrieval allow us to conclude that the proposed approach based on multidimensional receptive field histograms and the distance from feature space $\epsilon^2(\mathbf{x})$ provides a means for the classification of images. By using the same technique for multidimensional receptive field histograms of object appearances or even histograms representing entire 3D objects may provide a powerful framework for the classification of objects. An important question herein is the automatic extraction of histograms which are representative for a visual class. This topic will be discussed briefly in the following section.

9.4 The extraction of visual classes

The preceding section has shown examples for the recognition of visual classes based on the distance from feature space introduced in section 9.2. The results underline the validity of the proposed algorithm for the recognition of visual classes.

In the examples of the preceding section, the sample images for a visual class have been chosen manually. This is neither desirable nor feasible in general. We therefore discuss here the extraction of histograms for a visual class. There are at least two complementary requirements for the extraction of histograms H_i of a visual class:

- the histograms H_i should correspond to visual similar object appearances
- the histograms H_i should be representative of the visual class so that we can compute the eigenvector representation and estimate the probability density $\hat{p}(\mathbf{x}|\Omega)$

For the purpose of illustration, we may take a closer look at the recognition of the visual class FLOWER of the preceding section. The three manually chosen flower images correspond to visually similar objects since all three images contain flowers. At the same time these three images correspond to the maximum distance of any triplet of the 7 flower images of figures 9.1(a) through (g). The maximal distance can be seen as a criteria to extract *representative* samples for the class flower. Overall, there exist 9 images of the 200 images, which may be classified flower by a human. If we, for example, add one of the 2 remaining flower images (for example the one with the smallest distance in feature, which guarantees that an independent dimension is added to the eigenspace), all flower images of the database are within the first matches.

In the following we define an algorithm for the extraction of histograms corresponding to visually similar objects. In order to obtain representative samples in general, more sample images are required than available in the experiments described above. Assuming a sufficient number of samples we can estimate the probability density function $\hat{p}(\mathbf{x}|\Omega)$ robustly. In such a case the recognition algorithm as defined in section 9.2 is expected to provide a reliable means for the recognition of visual classes. In the previous section we did not assume a sufficient number of samples but a small number of manually chosen examples. Such a context is typical in the case of image retrieval where a human may choose samples and counter samples for a image database query [Pen 96]. How to obtain a representative set of images in the general case is the topic of future research.

In the following we define a clustering algorithm which may extract histograms from a database corresponding to visually similar objects.

Clustering based on intersection

Here we propose a clustering algorithm in the context of a histogram database. The clustering algorithm, as many classical clustering algorithms, is based on two steps:

- find the most similar histogram pair in the histogram database
- merge this pair to a representation of both histograms

Different measures may be used to determine the most similar histogram pair. In principal, any of the measurements proposed in section 5.1 can be used. The most successful comparison measurement for recognition has been the χ^2–statistics χ^2_{qv}. Unfortunately this measurement is not a metric which makes it inappropriate in the context of clustering. We therefore propose to use the intersection measurement $\cap$, which provides the overall next best recognition results. In particular, the intersection measurement also performed well in the presence of partial occlusion.

The application of the intersection measurement $\cap$ for the comparison of histograms allows the merging of two histograms based on an intersection operation. We define the merging operation of two histograms $Q = \cup_{\mathbf{i}} q_{\mathbf{i}}$ and $V = \cup_{\mathbf{i}} v_{\mathbf{i}}$ as follows:

$$merge(Q, V) = \cup_{\mathbf{i}} \cap (q_{\mathbf{i}}, v_{\mathbf{i}}) = \cup_{\mathbf{i}} \min(q_{\mathbf{i}}, v_{\mathbf{i}}) \tag{9.135}$$

This merging operation allows the extraction of the common part of two histograms. Such common parts of histograms may correspond to visually similar parts of the corresponding images. In order to emphasize this statement we applied the clustering algorithm defined here onto the database of the 200 countryside images used in the preceding section.

Figure 9.3 shows an example of a cluster, obtained by the described algorithm. The white squares signify images which have been clustered and merged with the defined merging operation. As we can see the clustered images are visually similar. This let us to conclude that we can extract histograms of a visual class based on the proposed clustering algorithm. As mentioned early it is also important to extract representative samples, which is the topic of future research.

9.5 Conclusion

This chapter has introduced the concept of visual classes as a general framework for object classification. Based on a statistical representation of objects the chapter develops the extraction

Figure9.3: Clustered images. Shown are coastal images which have been clustered automatically within a database of 200 countryside images

and the representation of visual classes. Visual classes are defined by 2D– and/or 3D–similarities of object appearances. Similarities in 2D correspond to visually similar appearances of objects. Similarities in 3D are much more restrictive, since the entire 3D shape and the texture on all surfaces have to be similar for both objects. Since visual classes are only based on image measurements they can be extracted reliably from images. Furthermore visual classes are context independent whereas object classes are in general context depended (in space and time). We therefore argue that the concept of visual classes provides a powerful framework for the classification of objects. The topic of this chapter therefore extends the application of multidimensional receptive field histograms to the recognition of visual classes as a subproblem of object classification.

The chapter proposes a maximum–likelihood algorithm for the recognition of visual classes. An application of the algorithm in the context of image retrieval shows the validity of the concept. Examples are given for example for the retrieval of coastal images from a database of 200 images. The important question of the extraction of histograms corresponding to a visual class is discussed. We define a clustering algorithm which is able to extract such histograms. Nevertheless, the question of how to extract histograms corresponding to a visual class that are representative for a visual class, will be topic of future research.

From a more general point of view this chapter proposes the concept of visual classes for *appearance based object classification*. As stated in the introduction this topic is a fundamental challenge for appearance based recognition schemes in general. As the concept of visual classes that has been proposed here it may fill the gap between appearance based object identification and a general object classifier.

Chapter 10

Conclusion and perspective

Being inspired by the robustness and speed of the color histogram approach of Swain and Ballard, our aim was to generalize their approach. In particular, we use statistics of arbitrary vectors of receptive field functions. From the generalization we derived a general statistical object representation framework. As a special representation of the statistical representation scheme we motivated and proposed multidimensional receptive field histograms. Different algorithms for the identification and classification of objects have been proposed and evaluated experimentally. In particular, we analyzed histogram matching, probabilistic object recognition and active object recognition. For the classification of objects we proposed a maximum–likelihood framework.

The guidelines for the study have been twofold: firstly, the *degrees of freedom of object recognition* listed in section 1.2 have been the guideline for the development of the statistical object representation and for the experimental evaluation of the proposed recognition algorithms. A second guideline has been the *challenges of appearance model* based approaches summarized in section 1.4.

10.1 Principal results

Here we want to list results and conclusions of the thesis, which have been learned. In particular, we may list three principal investigations and results of the thesis as follows:

- vectors of local characteristics and, in particular, the statistics of these vectors provide a reliable means for the representation and recognition of objects

- the probabilistic object recognition algorithm defined in chapter 7 is capable of recognizing objects without relying on any correspondence

- an appearance based classification scheme of objects has been proposed and applied in the context of content based image retrieval

All three point are of general interest for computer vision research. The first point emphasizes that object representation schemes based on vectors of local characteristics are highly discriminant and robust. This result can be used directly by other researchers. The second point is noteworthy, since almost all object recognition algorithms rely on correspondence between the testimage and objects of the database. Since it is generally difficult and computational expensive to obtain correspondence, the proposed probabilistic algorithm allows the computation of object hypothesis. The third point concerns a fundamental criticism of appearance based models. Chapter 9 defined a framework for appearance based classification and provides evidence that such a framework could work for the classification of objects. Therefore, the perspective of the thesis is the extension to appearance based object classification.

Additionally to the three points listed above we like to emphasize the following results of the thesis:

- development of an analogy between object recognition and information theory. This analogy allows the evaluation of the set of local neighborhood operators in the context of a certain recognition task

- the active object recognition algorithm defined in chapter 8 provided 3D–consistent object recognition based on several 2D views of an object. That is to say that the 3D information, inherent in the set of multidimensional receptive field histograms of an object, can be used implicitly without extracting explicitly 3D–information from the image

- the initial color histogram matching approach has been generalized to the matching of multidimensional receptive field histograms. Several histogram comparison measurements have been proposed, analyzed and applied for the recognition of 103 objects in the presence of image rotation, scale changes and viewpoint changes

10.2 Perspectives

Chapters 8 and 9 have outlined two major directions of perspectives of the thesis. These two and other possible directions of future research will be discussed here:

Active object recognition Chapter 8 introduced concepts and described preliminary results for the active recognition of objects. Due to lack of an appropriate active camera head, we could not perform further experiments. Therefore, it will be interesting to apply the developed principles for active exploration by a robot or in the context of human–computer interaction for recognition and tracking.

Appearance based classification The application of appearance based classification are manyfold. Among the most prominent candidates are content based image retrieval, automatic object modeling and generalization from examples. The propositions in chapter 9 may be seen as the first attempt to provide an algorithm for appearance based object classification. Different learning techniques may be examined in order to obtain a general framework for the extraction of visual classes from a representation based on vectors of local characteristics.

Extension to other local characteristics Most experiments described in the thesis are based on vectors of Gaussian derivatives employed as local descriptors for objects. As a direct extension of the approach we will apply descriptors based on color invariance introduced in section 3.1.3. However, such vector of local characteristics are inappropriate for the representation of objects like polyhedra with plain surfaces. So it is natural to exploit the technique using other types of local characteristics as for example moments, invariance descriptor and contour descriptor.

The extension to other local characteristics may provide object hypothesis derived from different types and sets of characteristics. The probabilistic framework allows us to incorporate such information coming from different sources in a convenient and elegant way. We can also imagine a hierarchy of characteristics where at each level of the hierarchy object hypothesis are computed on the basis of an appropriate set of characteristics.

Clustering of representations The clustering of multidimensional receptive field histograms offers the possibility of reducing the memory consumption per object. This is interesting in the case of large object databases. Also in the case of content based image retrieval clustering would enable an access hierarchy for images, reducing both memory and execution time. Chapter 9 uses the Karhunen–Loeve transformation for the compact representation of histograms. This or other dimensionality reduction schemes may be used also for the purpose of memory reduction.

Multiple sensor recognition As mentioned in the introduction, the statistical framework allows the incorporation of information coming from different sources. Such sources are a priori knowledge, information coming from different sets of local characteristics and from different sensors as camera and laser range finder. The statistical framework allows us to incorporate the information coming from different sources.

Appendix A

The image databases

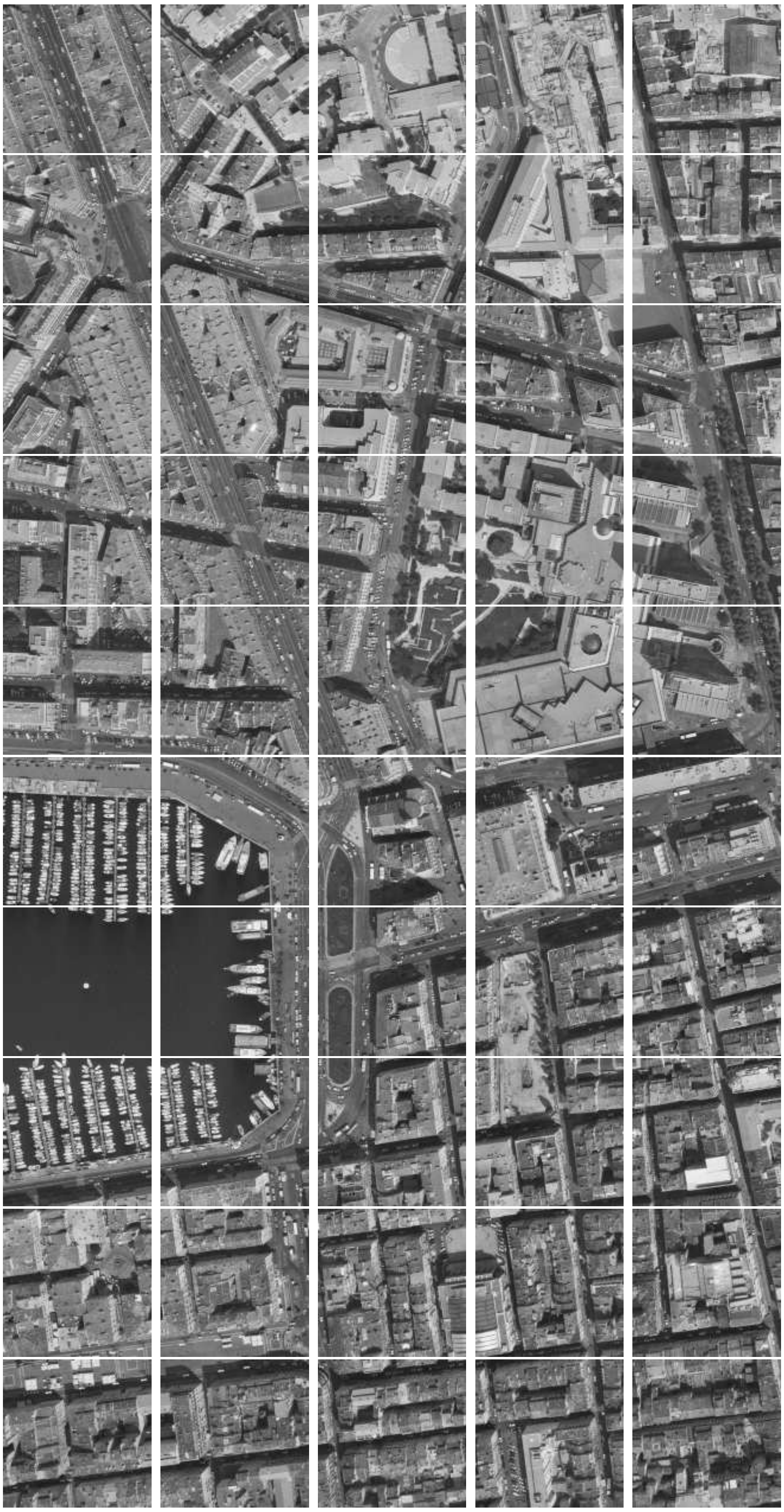
The following image databases have been used in different chapters:

1. 30 objects at six scales: figure A.1
2. 22 objects at 18 image plane rotations and three scales: figure A.2
3. 31 additional objects: figure A.3
4. 100 aerial images of Marseille, provided by the French company ISTAR. We also like to thank Roger Mohr who made this possible: figures A.4 and A.5
5. Columbia image database of 100 objects: figure A.6

FigureA.1: 30 × 6 scale images: for each of the objects we recorded six images at different scales. The scale factor between the first ant the last image is approximately 2.1. Figure 6.4 shows the six images for two objects

FigureA.2: $22 \times 18 \times 3$ *image plane rotation and scale images: for 22 objects we recorded 18 images corresponding to different image plane rotations. For each of these images we recorded three images corresponding to a scale change of approximately* ± 1.75*. The three scales for one object at a particular image plane rotation are shown in figure 6.5*

FigureA.3: 31 × 2 images of 31 objects. The two images have been recorded with slight changes in illumination, focus, orientation and scale



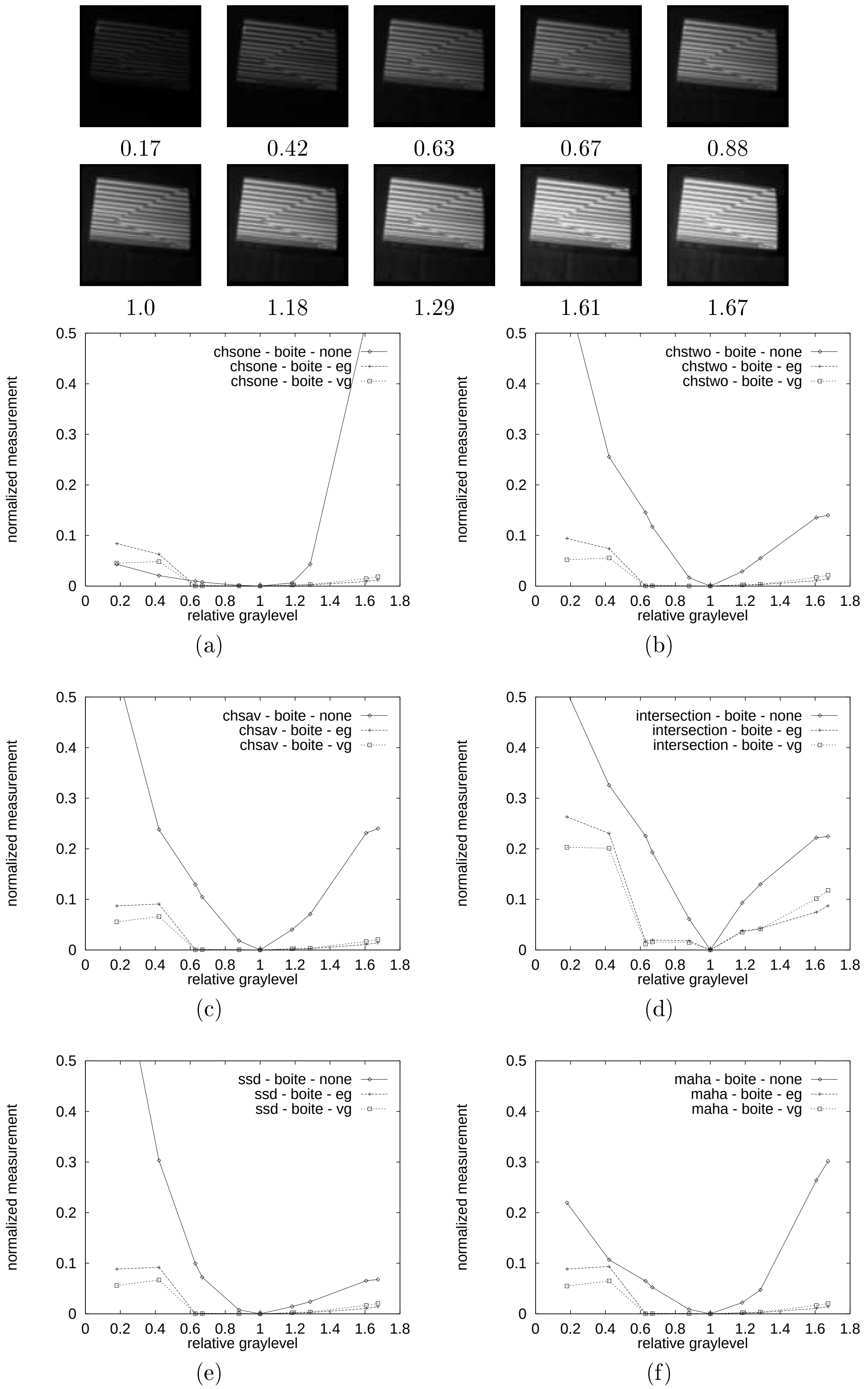
FigureA.4: The first half of 100 aerial images of Marseille, provided by ISTAR, France. There exist three further series of 100 aerial images of the same region of Marseille, corresponding to view point changes, caused by the movement of the airplane

FigureA.5: The second half of 100 aerial images of Marseille, provided by ISTAR, France. See for comments figure A.4

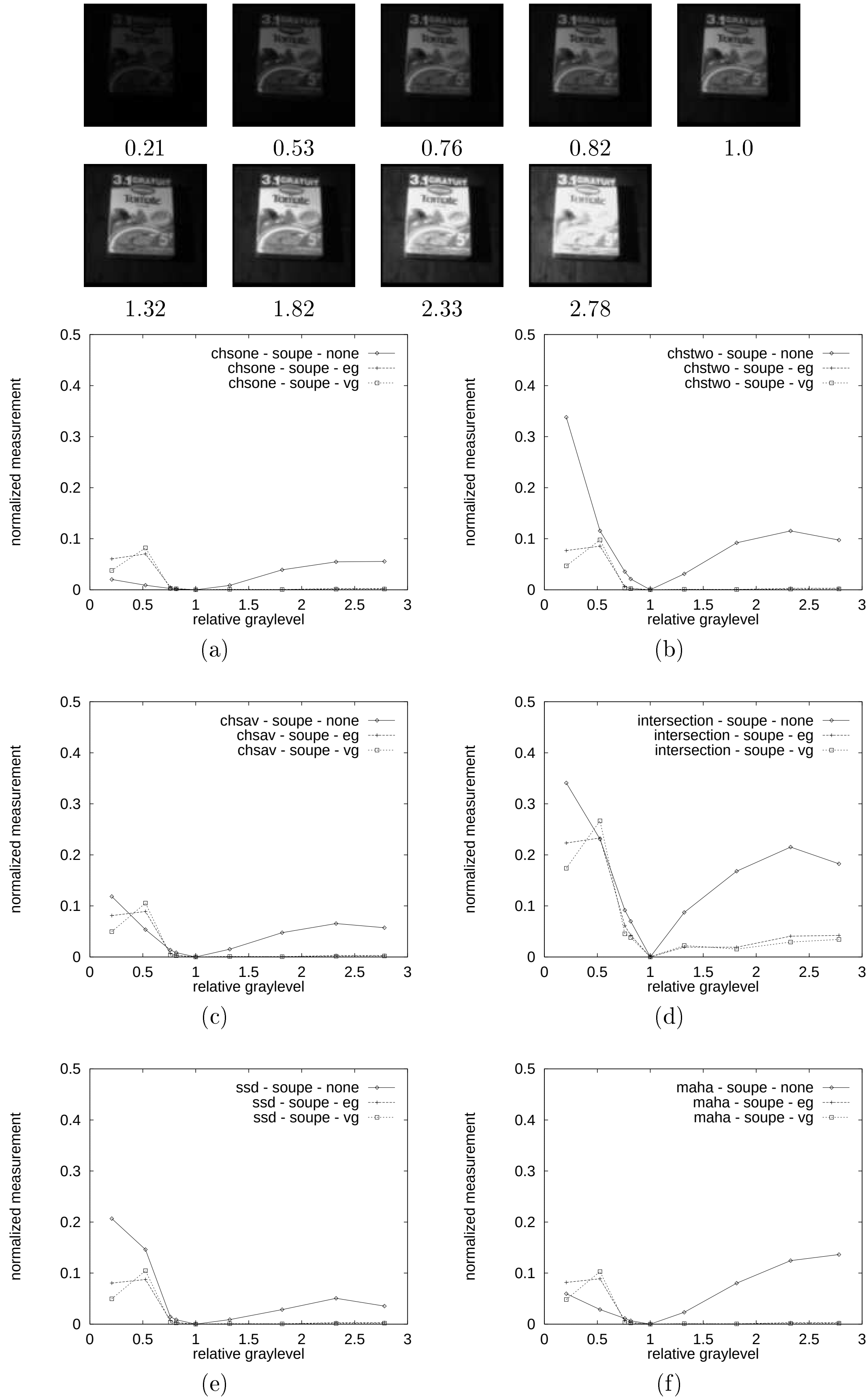
FigureA.6: 100 × 72 images for 100 objects of the Columbia image database. The 72 images correspond to different viewpoints 5° apart.

Appendix B

Collection of additional results



FigureB.1: Light intensity changes for the series "boite" and the stability of different normalization techniques and different histogram comparison measurements



FigureB.2: Light intensity changes for the series "soupe" and the stability of different normalization techniques and different histogram comparison measurements

B.1 Additional results for chapter 9: Object classification

(a) *Prototype*

(b) *Prototype*

(c) *Prototype*

(d) *First match*

(e) *Second match*

(f) *Third match*

(g) *Fourth match*

(h) *Fifth match*

FigureB.3: Forest recognition

(a) *Prototype* (b) *Prototype* (c) *Prototype*

(d) *First match* (e) *Second match* (f) *Third match*

(g) *Fourth match* (h) *Fifth match*

FigureB.4: Glacier recognition

(a) *Prototype*

(b) *Prototype*

(c) *Prototype*

(d) *First match*

(e) *Second match*

(f) *Third match*

(g) *Fourth match*

(h) *Fifth match*

FigureB.5: Waterfall recognition

List of Figures

List of Tables

Bibliography

[Bal 82] D.H. Ballard and C.M. Brown. *Computer Vision.* Prentice Hall, 1982.

[Bal 93] D.H. Ballard and L.E. Wixson. Object recognition using steerable filters at multiple scales. In *IEEE Workshop on Qualitative Vision*, pages 2–10, June 1993.

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

[Bar 96] K. Barnard, G. Finlayson, and B. Funt. Colour constancy for scences with varying illumination. In *ECCV'96 Fourth European Conference on Computer Vision, Volume II*, pages 3–15, April 1996.

[Beb 95] G. Bebis, M. Goergiopoulos, and N. da Vitoria Lobo. Learning geometric hashing functions for model–based object recognition. In *ICCV'95 Fifth International Conference on Computer Vision*, pages 543–548, 1995.

[Bei 94] J.S. Beis and D.G. Lowe. Learning indexing functions for 3-d model based object recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 275–280, 1994.

[Bel 89] Z.W. Bell. A bayesian/monte carlo segmentation method for images dominated by gaussian noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(9): 985–990, September 1989.

[Bel 96] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegmann. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In *ECCV'96 Fourth European Conference on Computer Vision, Volume I*, pages 45–58, 1996.

[Bel 97] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Recognition of images in large databases using color and texture. submitted to *International Conference on Computer Vision and Pattern Recognition*, 1997.

[Bla 96] M.J. Black and A.D. Jepson. Eigentracking: robust matching and tracking of articulated objects using a view–based representation. In *ECCV'96 Fourth European Conference on Computer Vision, Volume I*, pages 329–342, 1996.

[Bob 95] P. Bobet. *Tête stéréoscopique, Réflexes oculaires et Vision*. PhD thesis, INPG France, 1995.

[Bre 93] T.M. Breuel. Higher–order statistics in object recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 707–708, 1993.

[Bur 90] J. Burns, R. Weiss, and E. Riseman. View variation of point set and line segment features. In *Proceedings DARPA Image Understanding Workshop*, pages 650–659, 1990.

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

[Cal 93] A. Califano and R. Mohan. Systematic design of indexing strategies for object recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 709–710, 1993.

[Can 86] J.F. Canny. Computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6): 679–698, June 1986.

[Cle 91] D.T. Clemens and D.W. Jacobs. 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, 1991.

[Col 96] V. Colin de Verdière. Reconnaissance d'Objets par leurs Statistiques de Couleurs. Master's thesis, GRAVIR, I.N.P. Grenoble, 1996. DEA d'Imagerie, Vision et Robotique, in French.

[Cro 96] J.L. Crowley. Multi–modal tracking for video compression. In *Proceedings of the 4th International Symposium in Intelligent Robotic Systems*, pages 317–324, July 1996.

[Dau 93] J.G. Daugman. High confidence visual recognition of persons by test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11): 1148–1161, November 1993.

[Der 93] R. Deriche. Recursively implementing the gaussian and its derivatives. Technical Report 1893, INRIA–Sophia Antipolis, April 1993.

[Ede 93] S. Edelman. On learning to recognize 3-d objects from examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(8): 833–837, 1993.

[Egg 93] D.W. Eggert, K.W. Bowyer, C.R. Dyer, H.I. Christensen, and D.B. Goldgof. The scale space aspect graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11): 1114–1130, 1993.

[Enn 93] F. Ennesser and G. Medioni. Finding waldo, or focus of attention using local color information. In *International Conference on Computer Vision and Pattern Recognition*, pages 711–712, 1993.

[Enn 95] F. Ennesser and G. Medioni. Finding waldo, or focus of attention using local color information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8): 805–809, 1995.

[Fau 92] O.D. Faugeras, J.L. Mundy, N. Ahuja, C.R. Dyer, A.P. Pentland, R. Jain, K. Ikeuchi, and K.W. Bowyer. Why aspect graphs are not (yet) practical for computer vision. *CVGIP*, 55(2): 212–218, 1992.

[Fin 93] G.D. Finlayson, M.S. Drew, and B.V. Funt. Diagonal transforms suffice for color constancy. In *ICCV'93 Fourth International Conference on Computer Vision*, pages 164–171, 1993.

[Fin 95a] G.D. Finlayson. Color constancy in diagonal chromatic space. In *ICCV'95 Fifth International Conference on Computer Vision*, pages 218–223, 1995.

[Fin 95b] G.D. Finlayson, B.V. Funt, and K. Barnard. Color constancy under varying illumination. In *ICCV'95 Fifth International Conference on Computer Vision*, pages 720–725, 1995.

[Fin 96] G.D. Finlayson, S.C. Chatterjee, and B.V. Funt. Color angular indexing. In *ECCV'96 Fourth European Conference on Computer Vision, Volume II*, pages 16–27, 1996.

[FL 95] G. Funka-Lea and R. Bajcsy. Combining color and geometry for the active, visual recognition of shadows. In *ICCV'95 Fifth International Conference on Computer Vision*, pages 203–209, 1995.

[Fli 95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Juang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, pages 23–32, September 1995.

[Flo 91] L.M.J. Florack, B.M. ter Haar Romeny, J.J. Koenderink, and M.A. Viergever. General intensity transformations and second order invariants. In *Proceedings of the 7th Scandinavian Conference on Image Analysis*, pages 338–345, 1991.

[For 90] D.A. Forsyth. A novel algorithm for color constancy. *International Journal of Computer Vision*, 5(1): 5–36, 1990.

[Fre 91] W.T. Freeman and E.H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9): 891–906, 1991.

[Fuk 90] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Computer Science and Scientific Computing. Academic Press, New York, 2nd edition, 1990.

[Fun 91] B.V. Funt, M.S. Drew, and J. Ho. Color constancy from mutual reflection. *International Journal of Computer Vision*, 6(1): 5–24, 1991.

[Fun 95] B.V. Funt and G.D. Finlayson. Color constant color indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5): 522–529, 1995.

[Gab 46] D. Gabor. Theory of communication. *Proc. Inst. Elec. Eng.*, 93(26): 429–441, 1946.

[Gor 94] M.M. Gorkani and R.W. Picard. Texture orientation for sorting photos "at a glance". In *ICPR'94 Proceedings of the International Conference on Pattern Recognition*, pages 259–464, October 1994.

[Gri 90] W.E.L. Grimson and D.P. Huttenlocher. On the sensitivity of the hough transform for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(3): 255–274, March 1990.

[Gri 91] W.E.L. Grimson and D. Huttenlocher, editors. Special issue on interpretation of 3-d scenes – part i. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10), October 1991.

[Gri 92] W.E.L. Grimson and D. Huttenlocher, editors. Special issue on interpretation of 3-d scenes – part ii. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2), February 1992.

[Gui 96] L.J. Guibas and C. Tomasi. Image retrieval and robot vision research at stanford. In *ARPA Image Understanding Workshop*, 1996.

[Haf 95] J. Hafner, H.S. Sawhney, W. Equitz, M. Flickner, and W. Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7): 729–736, July 1995.

[Hea 94] G. Healey and D. Slater. Using illumination invariant color histogram descriptors for recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 355–360, 1994.

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

[Hor 96] J. Hornegger. *Statistische Modellierung, Klassifikation und Lokalisierung von Objekten.* Shaker Verlag GmbH, 1996. PhD thesis, in German, see also [Hor 95].

[Hun 94] M. Hunke. Locating and tracking of human faces with neural networks. Technical Report CMU–CS–94–155, Carnegie Mellon University, August 1994.

[Hut 87] D.P. Huttenlocher and S. Ullman. Object recognition using alignment. In *ICCV'87 First International Conference on Computer Vision*, 1987.

[Hut 95] D. Huttenlocher and E.W. Jaquith. Computing visual correspondence: Incorporating the probability of a false match. In *ICCV'95 Fifth International Conference on Computer Vision*, pages 515–520, 1995.

[Int 93] N. Intrator and J.I. Gold. Three–dimensional object recognition using an usupervised bcm network: The usefulness of distinguishing features. *Neural Computation*, 5: 61–74, 1993.

[Jon 92] D.G. Jones and J. Malik. 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, 1992.

[Kan 78] T. Kanade. Region segmentation: Signal vs. semantics. In *Proceedings of the 4th International Conference on Pattern Recognition*, pages 95–105, 1978.

[Kir 90] M. Kirby and L. Sirovich. Application of the karhunen–loève procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1): 103–108, January 1990.

[Koe 79] J.J. Koenderink and A.J. Doorn. The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32: 211–216, 1979.

[Koe 84] J.J. Koenderink. The structure of image. *Biological Cybernetics*, 50: 363–396, 1984.

[Koe 87] J.J. Koenderink and A.J. Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55: 367–375, 1987.

[Koe 91] J.J. Koenderink. Local image structure. In *Proceedings of the 7th Scandinavian Conference on Image Analysis*, pages 1–7, 1991.

[Lam 88a] Y. Lamdan, J.T. Schwartz, and H.J. Wolfson. Object recognition by affine invariant matching. In *International Conference on Computer Vision and Pattern Recognition*, pages 335–344, 1988.

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

[Lam 88c] Y. Lamdan and H.J. Wolfson. On recognition of 3-d objects from 2-d images. In *International Conference on Robotics and Automation*, pages 1407–1413, 1988.

[Lam 96] B. Lamiroy and P. Gros. Rapid object indexing and recognition using enhanced geometric hashing. In *ECCV'96 Fourth European Conference on Computer Vision, Volume I*, pages 59–70, 1996.

[Mal 85] L.T. Maloney. *Computional approaches to color constancy.* PhD thesis, Stanford Unversity, 1985.

[Mal 86] L.T. Maloney and B.A. Wandell. Color constancy: A method for recovering surface spectral reflectance. *Journal of the Optical Society of America A*, 3(1): 29–33, January 1986. Interest: Color Constancy, CV, Co.

[Mal 89] J. Malik and P. Perona. A computational model of texture segmentation. In *International Conference on Computer Vision and Pattern Recognition*, pages 326–332, 1989.

[Mar 78] D. Marr. Representing visual information – a computational approach. In A.R. Hanson and E.M. Riseman, editors, *Computer Vision Systems*, pages 61–80. Academic Press New York, 1978.

[Mar 97] J. Martin and J.L. Crowley. An appearance-based approach to gesture-recognition. to appear in ICIAP'97, Florence, Italy, September 1997.

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

[Mel 96] B.W. Mel. SEEMORE: Combining color shape, and texture histogramming in a neurally-inspired approach to visual object recognition. Technical report, Department of Biomedical Engineering, University of Southern California, MC 1451, 1996. appeared in *Proceedings of International Conference of Pattern Recognition 1996*.

[Mog 95] B. Moghaddam and A. Pentland. Maximum likelihood detection of faces and hands. In *International Workshop on Automatic Face– and Gesture–Recognition*, pages 122–128, 1995.

[Moh 97] R. Mohr, S. Picard, and C. Schmid. Bayesian decision versus voting for image retrieval. submitted to CAIP'97, Kiel, Germany, 1997.

[Mun 92] J. L. Mundy and Andrew Zisserman, editors. *Geometric Invariance in Computer Vision*. MIT Press, 1992.

[Mun 93] J. L. Mundy, A. Zisserman, and D. Forsyth, editors. *Application of Invariance in Computer Vision*. Number 825 in Lecture Notes in Computer Science. Springer Verlag, 1993.

[Mur 93] H. Murase and S. K. Nayar. Learning and recognition of 3-d objects form brightness images. In *AAAI Fall Symposium, Machine Learning in Computer Vision*, pages 25–29, October 1993.

[Mur 95] H. Murase and S.K. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, 14: 5–24, 1995.

[Nag 92] H.-H. Nagel. Reflections on active (machine) vision. In A.K. Sood and H. Wechsler, editors, *Active Perception and Robot Vision*, pages 23–42. Springer Verlag, 1992.

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

[Nel 95] R.C. Nelson. 3-d recognition via 2-stage associative memory. Technical Report TR 565, Department of Computer Science, University of Rochester, January 1995.

[Nel 96] R.C. Nelson. Memory-based recognition for 3-d objects. In *ARPA, Image Understanding Workshop*, pages 1305–1310, February 1996.

[Nen 96] S.A. Nene, S.K. Nayar, and H. Murase. Columbia object image library (coil-100). Technical Report CUCS-006-96, Department of Computer Science, Columbia University, 1996.

[Obj 96] International workshop on object representation for computer vision. Cambridge, England, April 1996.

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

[Pen 94] A. Pentland, B. Moghaddam, and T. Starner. View–based and modular eigenspaces for face recognition. In *International Conference on Computer Vision and Pattern Recognition*, 1994. see also M.I.T. Media Laboratory TR 245.

[Pen 96] A. Pentland, R.W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3): 233–254, 1996.

[Per 95] P. Perona. Deformable kernels in early vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5): 488–499, May 1995.

[Pic 96] R.W. Picard. A society of models for video and image libraries. Technical Report 360, M.I.T. Media Laboratory, 1996. submitted to IBM Systems Journal.

[Pog 90] T. Poggio and S. Edelman. A network that learns to recognize 3d objects. *Nature*, 343, January 1990.

[Pop 93] A.R. Pope and D.G. Lowe. Learning object recognition models from images. In *ICCV'93 Fourth International Conference on Computer Vision*, pages 296–301, 1993.

[Pop 94] K. Popat and R.W. Picard. Cluster–based probability model applied to image restoration and compression. In *IEEE Conference on Acoustics, Speech and Signal Processing*, Adeline, Australia, April 1994. also M.I.T. Media Laboratory TR 253.

[Pop 95] A.R. Pope. *Learning to Recognize Objects in Images: Acquiring and Using Probabilistic Models of Appearance.* PhD thesis, Department of Computer Science, University of British Columbia, 1995.

[Pop 96] A.R. Pope and D.G. Lowe. Learning appearance models for object recognition. In *International Workshop on Object Representation for Computer Vision*, Cambridge, England, April 1996.

[Pre 92] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C.* Cambridge University Press, 2nd edition, 1992.

[Rao 95a] R.P.N. Rao and D.H. Ballard. An active vision architecture based on iconic representations. *Artificial Intelligence (Special Issue on Vision)*, 78: 461–505, 1995.

[Rao 95b] R.P.N. Rao and D.H. Ballard. Object indexing using an iconic sparse distributed memory. In *ICCV'95 Fifth International Conference on Computer Vision*, pages 24–31, 1995.

[Rez 95] F.M. Reza. *An Introduction to Information Theory.* Dover Publications, New York, 1995.

[Rig 93] I. Rigoutsos and R. Hummel. Distributed bayesian object recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 180–186, 1993.

[Rom 91] B.M. ter Haar Romeny, L.K.J. Florack, J.J. Koenderink, and M.A. Viergever. Invariant third order properties of isophotes: T-junction detection. In *Proceedings of the 7th Scandinavian Conference on Image Analysis*, pages 346–353, 1991.

[Ros 84] A. Rosenfeld. Image analysis: Problems progress and prospects. *Pattern Recognition*, 17(1): 3–12, 1984.

[Sch 95a] B. Schiele and J.L. Crowley. Object recognition using multidimensional receptive field histograms and its robustness to view point changes. Presented at *1995 Rosenon Workshop on Computational Vision*, July 1995.

[Sch 95b] B. Schiele and A. Waibel. Estimation of the head orientation based on a face–color–intensifier. In *3rd International Symposium on Intelligent Robotic Systems '95*, pages 339–346, 10–14 July 1995.

[Sch 95c] B. Schiele and A. Waibel. Gaze–tracking based on face–color. In *IWAFGR 95, International Workshop on Automatic Face–and Gesture–Recognition*, pages 344–349, June 1995.

[Sch 96a] B. Schiele and J.L. Crowley. Object recognition using multidimensional receptive field histograms. In *ECCV'96 Fourth European Conference on Computer Vision, Volume I*, pages 610–619, 14–16 April 1996.

[Sch 96b] B. Schiele and J.L. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. In *ICPR'96 Proceedings of the 13th International Conference on Pattern Recognition, Volume B*, pages 50–54, August 1996.

[Sch 96c] B. Schiele and J.L. Crowley. The robustness of object recognition to rotation using multidimensional receptive field histograms. submitted. available via www[1], 1996.

[Sch 96d] B. Schiele and J.L. Crowley. Where to look next and what to look for. In *4th International Symposium on Intelligent Robotic Systems'96*, pages 139–146, July 1996.

[Sch 96e] B. Schiele and J.L. Crowley. Where to look next and what to look for. In *IROS'96 Intelligent Robots and Systems*, pages 1249–1255, Osaka, Japan, November 1996.

[Sch 96f] B. Schiele and J.L. Crowley. The robustness of object recognition to view point changes using multidimensional receptive field histograms. Presented at *ECIS-VAP meeting, Object Recognition Day.* Available via WWW[2], March 1996.

[1] http://pandora.imag.fr/Prima/schiele/

[2] http://pandora.imag.fr/Prima/schiele/

[Sch 96g] B. Schiele and J.L. Crowley. Reconnaissance des objets utilisant des histogrammes multidimensionnels des champs réceptive. In *Journée Orasis'96, Clermont–Ferrand, France*, pages 7–12, May 1996.

[Sch 96h] C. Schmid. *Appariement d'images par invariants locaux de niveaux de gris.* PhD thesis, I.N.P.Grenoble, 1996.

[Sch 96i] C. Schmid and R. Mohr. Combining grayvalue invariants with local constraints for object recognition. In *International Conference on Computer Vision and Pattern Recognition*, 1996.

[Sch 97a] B. Schiele and J.L. Crowley. The concept of visual classes for object classification. In *SCIA'97, Proceedings of the Scandinavian Conference on Image Analysis*, pages 43–50, June 1997.

[Sch 97b] B. Schiele and J.L. Crowley. Transinformation of object recognition and its application to viewpoint planning. *Robotics and Autonomous Systems*, 1997. To appear.

[Sch 98] B. Schiele and J.L. Crowley. Transinformation for active object recognition. In *ICCV'98 Sixth International Conference on Computer Vision*, January 1998. accepted.

[Shv 90] H. Shvaytser. Learnable and nonlearnable visual concepts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5): 459–466, 1990.

[Sir 87] L. Sirovich and M. Kirby. Low–dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America*, 4(3): 519–524, March 1987.

[Sla 95] D. Slater and G. Healey. 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, 1995.

[Str 92] M.A. Stricker. Color and geometry as cues for indexing. Technical Report CS 92–22, University of Chicago, November 1992.

[Str 94] M. Stricker and M. Swain. The capacity of color histogram indexing. In *International Conference on Computer Vision and Pattern Recognition*, pages 704–708, 1994.

[Swa 90] M.J. Swain. Color indexing. Technical Report 360, University of Rochester Computer Science, November 1990.

[Swa 91] M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1): 11–32, 1991.

[Swa 93a] M.J. Swain. Interactive indexing into image databases. In *SPIE, Storage and Retrieval for Image and Video Databases*, 1993.

[Swa 93b] M.J. Swain and M.A. Stricker. Promising directions in active vision. *International Journal of Computer Vision*, 11(2): 109–126, 1993.

[Tom 94] C. Tomasi and L.J. Guibas. Image descriptors for browsing and retrieval. In *ARPA Image Understanding Workshop*, November 1994.

[Tur 91a] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1): 71–86, 1991.

[Tur 91b] M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–591, June 1991.

[Wei 94] C.F.R. Weiman. Efficient discrete gabor functions for robot vision. In *SPIE Vol. 2242 Wavelet Applications*, pages 148–160, April 1994.

[Wes 92] C.-J. Westelius. *Preattentive Gaze Control for Robot Vision*. PhD thesis, Department of Electrical Engineering, Linköping University, 1992.

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

[You 86] R.A. Young. Simulation of human retinal function with the gaussian derivative model. In *International Conference on Computer Vision and Pattern Recognition*, pages 564–569, 1986.