
A computational recognition system grounded in perceptual research

Dissertation
zur Erlangung des Grades eines Doktors
der Naturwissenschaften
der Fakultät für Mathematik und Physik
der Eberhard-Karls-Universität Tübingen
vorgelegt von

Christian Wallraven

aus Kempen
2007

Tag der mündlichen Prüfung: 18. Oktober 2006

Dekan: Prof. Dr. N. Schopohl

1. Berichterstatter: Prof. Dr. H. Ruder / Prof. Dr. H. Büthoff

2. Berichterstatter: Prof. Dr. B. Schölkopf

3. Berichterstatter: Prof. Dr. W. Straßer

Contents

1	Cognitive basis of object recognition	1
1.1	Introduction	1
1.2	Cognitive psychophysics	4
1.2.1	Structural versus view-based approaches	4
1.2.2	View-based recognition of faces	6
1.2.3	The canonical view	15
1.2.4	Temporal aspects of object learning	18
1.2.5	Temporal aspects of object recognition	21
1.2.6	Configuration and components	26
1.3	Physiology	32
1.3.1	Visual processing in the brain	32
1.3.2	Beyond the traditional view	35
1.4	Conclusion	36
2	Computational approaches to object recognition	39
2.1	Data representation	40
2.1.1	Structured shape models	40
2.1.2	Statistical appearance models	43
2.2	Classification algorithms	45
2.2.1	K-means with n-nearest neighbor	46
2.2.2	Radial basis function networks	47
2.2.3	Support vector machines	47
2.3	Conclusion	49
3	A generic framework for object learning and recognition	51
3.1	Introduction	51
3.2	Learning and recognizing objects using keyframes	53
3.2.1	Related concepts	54
3.2.2	What defines a keyframe?	55
3.3	Discussion of the framework	55
3.3.1	Keyframes	56
3.3.2	Local visual features	57
3.4	Computational implementation	58
3.4.1	Visual features	58
3.4.2	Matching of visual features	59
3.4.3	Recognition and Incremental Learning	61
3.5	Conclusion	62

4	Cognitive modeling studies	63
4.1	View-based recognition of faces	63
4.1.1	Feature matching - the horizontal prior	64
4.1.2	Modeling psychophysical experiments	65
4.1.3	Summary	69
4.2	Configuration and components	70
4.2.1	The face representation	70
4.2.2	Feature matching	73
4.2.3	Modeling psychophysical experiments	74
4.2.4	Summary	79
4.3	Temporal aspects of recognition	80
4.3.1	Modeling temporal contiguity by learning keyframes	80
4.3.2	The influence of morphing on feature tracking	81
4.3.3	Learning keyframes from morphed or scrambled sequences	84
5	Computational studies I - Keyframes	93
5.1	Geometric constraints for local feature matching	93
5.1.1	Geometric constraints	94
5.1.2	Recognition under large view rotations	94
5.2	Keyframe extraction for learning of object representations	97
5.2.1	Parameters of keyframe extraction	97
5.2.2	Real-world sequences	99
5.2.3	Recognition using keyframes	99
5.3	Incremental build-up of object representations	105
5.3.1	Parameters of incremental learning	106
5.4	Conclusion	108
6	Computational studies II - SVMs and local features	109
6.1	Introduction	109
6.2	Support Vector Machines and local features	110
6.3	Local kernels	112
6.4	Experiments	114
6.4.1	Experimental Setup	115
6.4.2	Experimental Results: View Generalization	117
6.4.3	Experimental Results: SIFT versus Local Kernels	119
6.4.4	Experimental Results: Recognition under Noise	119
6.4.5	Experimental Results: Recognition using position constraints	120
6.5	Conclusion	120
7	Computational studies III - SVMs and keyframes	121
7.1	Algorithmic Overview	122
7.1.1	Image Sequence Processing	122
7.1.2	Feature Matching	124
7.1.3	Image and Feature Matching for Kernel Machines	125
7.1.4	Multi-class SVMs	126
7.2	Computational Experiments	127
7.2.1	Database and Representation	127
7.2.2	Classification of Images	128
7.2.3	Classification of Features	129
7.2.4	Experimental validation of positive definiteness	131
7.3	Conclusion	132

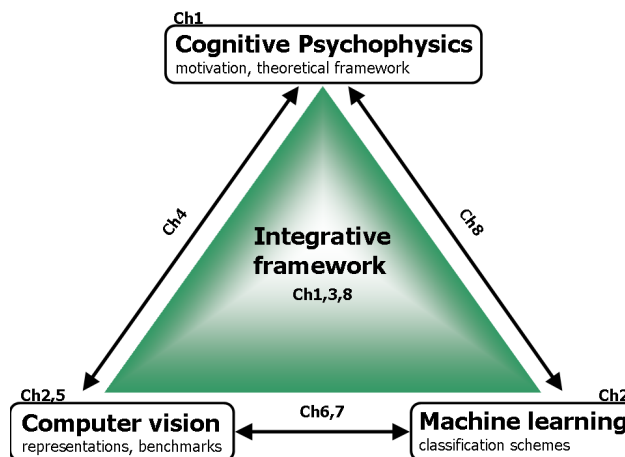
8	General conclusion and outlook	133
8.1	A unified framework for object recognition	134
8.1.1	Categorization processing by feature correspondences	134
8.1.2	The role of context in object recognition	136
8.1.3	Summary	136
8.2	Multi-modal keyframes	138
8.2.1	Psychophysics of visuo-haptic object recognition	138
8.2.2	Multi-modal keyframes - the view transition map	139
8.2.3	Computational experiments	141
8.3	Categorization using SVMs and local features	146
8.3.1	Experiment 1 - Categorization using a controlled database	146
8.3.2	Categorization experiments in cluttered scenes	146
8.3.3	Summary	149
	Bibliography	151

Summary

In this thesis a computational framework for visual object recognition is developed, which is based on results from perceptual research. The motivation for this approach is given by the fact that despite several decades of research in the field of computer vision, there still exists no recognition system which is able to match the visual performance of humans (or other primates). The apparent ease with which visual tasks such as recognition and categorization are solved by humans is testimony of a highly optimized visual system which not only exhibits excellent robustness and generalization capabilities but is in addition highly flexible in learning and organizing new data. In developing the framework, the underlying philosophy was to model object recognition on an abstract cognitive level rather than supplying a complete neurophysiologically plausible implementation. The proposed framework is able to model results from psychophysics and, in addition, delivers excellent recognition performance in computational recognition experiments. Furthermore, the framework also interfaces well with advanced classification schemes from machine learning thus further broadening the scope of application.

The basic outline of this thesis is summarized again in Figure 1, which situates this thesis in the context of the fields of cognitive psychophysics, computer vision as well as machine learning. As will be shown throughout the following chapters, the combination of methodologies from each field leads to an integrative framework, which has the potential to solve some of the open problems pertaining to object recognition.

Figure 1: Structure of this thesis.



The first main part of this thesis (chapter 1) reviews relevant results from both psychophysical and physiological studies on object recognition. A particular focus is placed on the *dynamic* aspect of recognition processes, the contribution of which has up to now been largely neglected in theoretical modeling of recognition. Several recent experiments in perceptual research have found the temporal dimension to play a large role in object recognition - both by being able to mediate learning of object representations and by providing an integral part of the representation

itself. In addition, results from further psychophysical studies, which were conducted in the scope of this thesis, shed light on how objects might be represented in the brain using local pictorial features and their spatial relations thus forming a sparse and at the same time structured object representation.

In chapter 2, an overview of current methodologies in computer vision and machine learning - two fields concerned with object recognition from a computational perspective - is given. In particular, methods from computer vision focusing on robust data representations and from machine learning focusing on efficient classification schemes will be discussed in a unified perspective.

Drawing on the three fields of cognitive research, computer vision and machine learning, chapter 3 develops an integrative and abstract framework for object recognition, which represents the core of this work. The main contribution of this integrative framework is that it provides spatio-temporal processing of visual input by means of a structured, appearance-based object representation.

The second main part of this thesis (chapter 4) is concerned with cognitive modeling of some of the recent psychophysical results presented in chapter 1 with the help of the proposed framework. Several experiments can successfully be modeled using a computational implementation of the framework which demonstrates the perceptual plausibility of spatio-temporal local feature processing. In addition, based on the computational modeling results, a number of performance predictions can be made, which can be tested in further psychophysical experiments thus closing the loop between experimental work and computational modeling.

The third part of the thesis (chapters 5-7) puts the proposed framework into a computational vision and machine learning perspective. Here, the main focus lies on approaches using *local features* and so-called *appearance-based* methods, that is, methods where images are represented by local pictorial features. Chapter 5 provides experimental validation of the integrative framework in a computer vision setting, which demonstrates the importance of using spatio-temporal information in several recognition experiments on both artificial and real-world data. In this context, another main contribution of this thesis consists of the development of a framework for combining spatio-temporal object representations based on local features and state-of-the-art statistical learning methods (Support Vector Machines, chapters 6-7). Within this novel framework, the proposed recognition approach is integrated and benchmarked against several other recognition algorithms. Extensive recognition experiments demonstrate that by combining efficient representations from computational vision with robust classification schemes from machine learning, excellent recognition performance can be achieved.

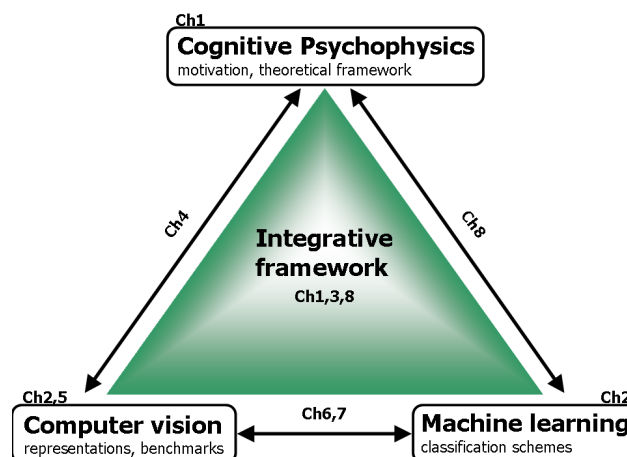
Finally, chapter 8 summarizes the application of the integrative framework in the previous chapters and presents an outlook for future work on combining cognitive research, computer vision and machine learning in order to provide solutions to some of the open problems in object recognition.

Zusammenfassung

In dieser Arbeit wird ein computergestütztes Framework für visuelle Objekterkennung entwickelt, das auf Ergebnissen aus der Wahrnehmungsforschung basiert. Die Motivation für diesen Ansatz leitet sich aus der Tatsache her, dass trotz mehrerer Jahrzehnte intensiver Forschung auf dem Gebiet des maschinellen Sehens (der "Computer Vision") noch immer kein Erkennungssystem existiert, das es mit der visuellen Erkennungsfähigkeit des Menschen aufnehmen kann. Die Leichtigkeit, mit der Erkennungs- und Kategorisierungsaufgaben vom Menschen gelöst werden, zeigt, dass wir ein hoch optimiertes visuelles System besitzen, das nicht nur robust erkennen und generalisieren kann, sondern auch flexibel neue Informationen lernen und organisieren kann. Bei der Entwicklung des hier vorgestellten Frameworks war die Philosophie, Objekterkennung auf einer abstrakten kognitiven Ebene zu modellieren, anstatt eine komplette neurophysiologisch plausible Implementation zu entwickeln. Das Framework ist in der Lage, Ergebnisse aus der Psychophysik zu modellieren und liefert zusätzlich sehr gute Erkennungsleistungen in Objekterkennungsexperimenten. Darüber hinaus ist es mit neuen, robusten Verfahren aus dem maschinellen Lernen ("Machine Learning") kombinierbar, was die Anwendungsmöglichkeiten noch erweitert.

Der Aufbau dieser Arbeit ist in Abb. 2 dargestellt und kombiniert kognitive Psychophysik, Computer Vision und Machine Learning. Wie in den folgenden Kapiteln gezeigt wird, hat die Kombination von Methoden und Ansätzen aus diesen Feldern in einem integrativen Framework das Potenzial, einige offene Probleme in der Objekterkennung anzugehen.

Figure 2: Aufbau dieser Arbeit.



Der erste Hauptteil dieser Arbeit (Kapitel 1) beginnt mit einem Überblick über relevante Forschung zur Objekterkennung aus den Gebieten der Psychophysik und der Physiologie. Ein besonderer Fokus liegt dabei auf dem *dynamischen* Aspekt von Erkennungsprozessen, deren Beitrag bisher bei der theoretischen Modellierung vernachlässigt wurde. Mehrere Experimente haben in letzter Zeit jedoch belegt, dass die zeitliche Dimension eine grosse Rolle bei der Objekterkennung spielt - dies sowohl beim Lernen von Objektrepräsentationen als auch als integraler Bestandteil der

Repräsentationen selber. Weitere Resultate aus Experimenten, die im Rahmen dieser Arbeit durchgeführt wurden, zeigen, wie Objekte im Gehirn durch lokale Merkmale und ihren raumzeitlichen Zusammenhang repräsentiert werden könnten, was zu sowohl kompakten als auch strukturierten Objektrepräsentationen führt. Kapitel 2 präsentiert einen Überblick aktueller Methoden in der Computer Vision und dem Machine Learning - zwei Felder, die sich mit Objekterkennung von der algorithmischen Seite her beschäftigen. Hier liegt der Fokus insbesondere auf der Integration von Methoden aus der Computer Vision, die sich mit robusten Datenrepräsentationen befassen, mit Methoden aus dem Machine Learning, die effiziente Klassifikationsalgorithmen bieten. Kapitel 3 entwickelt aus den drei Feldern der kognitiven Wahrnehmungsforschung, der Computer Vision und dem Machine Learning ein integratives, abstraktes Framework für Objekterkennung, das den Kern dieser Arbeit darstellt. Der Hauptbeitrag dieses integrativen Frameworks ist dabei ein System zur robusten Objekterkennung basierend auf strukturierten, bildbasierten Objektrepräsentationen, die räumliche und zeitliche Information integrieren.

Der zweite Hauptteil dieser Arbeit (Kapitel 4) beschäftigt sich mit der Modellierung einiger neuerer Psychophysikexperimente aus Kapitel 1 mithilfe des vorgeschlagenen Frameworks. Es wird gezeigt, dass eine Implementation des Frameworks in der Lage ist, die Ergebnisse der Experimente korrekt zu modellieren. Zusätzlich macht das Framework konkrete Voraussagen, die in weiteren psychophysischen Experimenten getestet werden können, womit sich der Kreis zwischen experimenteller und theoretischer Arbeit schliesst.

Der dritte Teil dieser Arbeit (Kapitel 5-7) stellt das vorgeschlagene Framework in den Kontext von Computer Vision und Machine Learning. Der Hauptfokus hier liegt auf der Evaluation von Methoden, die lokale Features in einem bildbasierten Ansatz benutzen. In Kapitel 5 wird das Framework mit Erkennungsexperimenten validiert, die die Wichtigkeit von raum-zeitlicher Information bei der Erkennung von Objekten in computergenerierten Sequenzen und Videosequenzen untermauern. Ein weiterer wichtiger Beitrag dieser Arbeit ist in diesem Kontext die Entwicklung einer Kombination von raum-zeitlichen Objektrepräsentationen auf Basis von lokalen Merkmalen und aktuellen Verfahren aus dem Machine Learning (den Support Vektor Maschinen). Kapitel 6-7 vergleichen diesen kombinierten Ansatz mit anderen Erkennungsalgorithmen und zeigen, dass die Kombination von effizienten Repräsentationen aus der Computer Vision mit robusten Klassifikationsalgorithmen aus dem Machine Learning eine exzellente Erkennungsleistung ermöglichen. Kapitel 8 schliesslich fasst die Arbeit zusammen und präsentiert einen Ausblick auf zukünftige Arbeit an der Schnittstelle von kognitiver Wahrnehmungsforschung, Computer Vision und Machine Learning, die helfen kann, offene Problem in der Objekterkennung anzugehen.

Chapter 1

Cognitive basis of object recognition

How the brain represents objects and scenes for the purpose of recognition is one of the most important questions in cognitive science. This chapter provides an overview of cognitive research on this question and focuses in particular on psychophysical studies on object and face recognition. The results from these studies are interpreted in the context of the most prominent theories of object recognition. Furthermore, several open questions pertaining to properties of recognition processes are addressed and analyzed in detail in psychophysical studies. In addition, a brief overview of relevant studies in cognitive neuroscience - more specifically in functional brain imaging techniques - is given. The chapter closes with a summary of several key issues that need to be addressed in order to build a computational recognition system firmly based on cognitive research.

1.1 Introduction

To illustrate the complexity and difficulty of an object recognition task consider Figure 1.1, which shows computer graphics renderings of a face of an individual under various viewing conditions. The images of this face under normal viewing conditions can exhibit large variations due to changes in pose, lighting or viewing distance (Figure 1.1 1-3). Another common variation is due to non-rigid deformations of the face which are caused by facial expressions (Figure 1.1 4). In addition, the object need not be fully visible because it is occluded by other objects, or leaves our field of view (Figure 1.1 5). Going even further, one could also include transformations which can be modeled by artificial image manipulations such as blurring, warping or image plane rotation. It is astonishing that humans nevertheless seem to be able to still recognize these faces correctly despite their large variations in appearance.

The previous example illustrated object recognition for the identification of a previously seen or learned object. The second class of recognition processes is called *categorization* and assigns category labels to familiar *and* unfamiliar visual objects. Categories in this context thus form a semantic container for visual objects according to some common properties - these properties are usually based on visual (red objects) and/or functional cues (something you can sit on)¹. To illustrate the process of categorization, Figure 1.2 shows a computer graphics generated picture of a room. When asked to point out all the *chairs* in the scene, one can do so easily despite large variations in size, illumination, visibility and shape of the chairs. The ease with which humans are able to categorize an object even though they have not had any previous experience with that particular exemplar is evidence of the striking robustness and generalization capabilities of visual recognition processes.

The third class of recognition processes deals with visual input on an even larger scale and

¹For a general introduction into categorization, refer to Rosch et al. [1976], Graf [2002], Graf et al. [2002]. See also chapter 8 for a more in-depth discussion of categorization processes.

Figure 1.1: Images of a face (o) under different viewing conditions including pose variation (1), direction of illumination (2), viewing distance (3), non-rigid deformation due to facial expression (4), occlusion (5), blurring (6), image warping (7) and plane rotation (8).

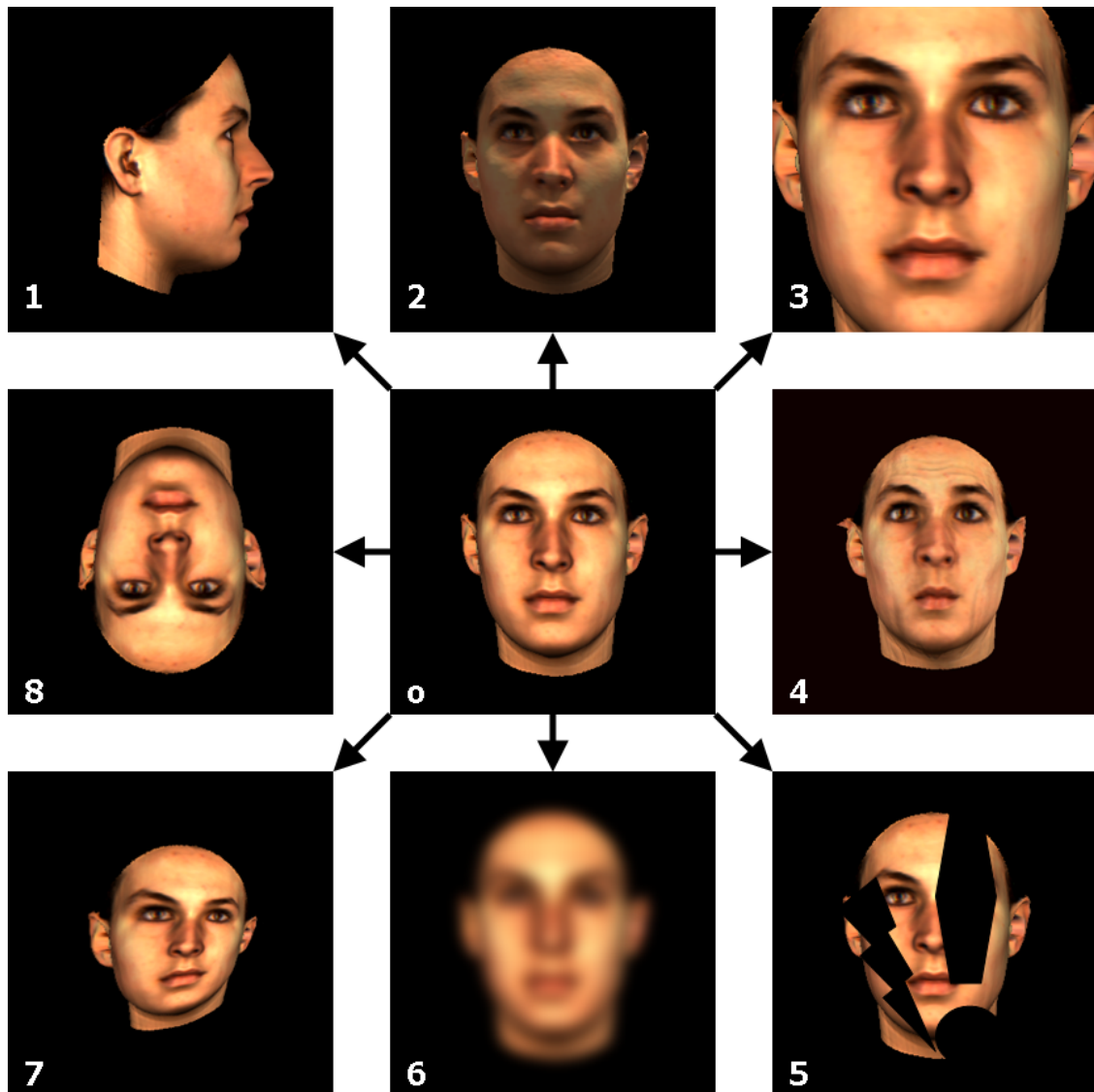


Figure 1.2: Computer generated image of an office room (from Bülthoff and Bülthoff [2003]) demonstrating the problems our visual system has to solve in categorizing every-day objects (count the chairs in this image!).



is involved in scene perception. Returning to Figure 1.2, the context of the scene as an indoor environment is immediately obvious to the observer. Further processing allows to refine this result to the context of an office room. This context and the resulting description of scene layout can for example be used when judging which chairs one could actually sit on - in this case the chair on the ceiling would be ruled out, and the chair on the desk would be classified as a small model of a chair even though its image size is exactly the same as that of the chair in the back room.

The above examples of the seemingly effortless and robust object recognition performance could lead to the conclusion that any object can be recognized from any and all viewpoints uniformly well. However, several researchers have shown that not all views of an object are equally good for object recognition tasks, even when the object is as familiar as a coffee mug. In other words, object recognition (and object categorization) performance is not view-invariant.

Another important question concerns the learning of objects, which in this case can be defined as the processes encoding the visual input to form an object representation suitable for efficient recognition: How does the brain actually know that different images of an object belong together? As an example, the views from an airplane from above and from the side form drastically different images on the retina but they are nevertheless reliably assigned to the category of "airplanes". In this thesis, I will show that an answer to this problem could lie in the *spatio-temporal* continuity of the visual input which helps to establish a coherent representation from sequences of images.

The evidence from this research also raises important issues for computational modeling of human object recognition and gives insights into an efficient and effective algorithmic implementation. Such biologically motivated vision systems could eventually achieve human performance levels, which is certainly one of the ultimate goals of computer vision. The aim of this thesis is thus to present results from cognitive research on human object recognition and then show a particular

instantiation of the findings in a computational system.

This chapter first gives a brief review of two current models for object recognition - a *view-based model* which predicts *view-dependent* recognition and a *structural model* which predicts *view-invariant* recognition. I will then discuss several psychophysical experiments investigating human object recognition performance, which use a range of tasks with unfamiliar objects (computer generated shapes) and highly familiar objects (for example, faces). The results from these experiments strongly favor the view-based approach to recognition. Furthermore, research on the temporal aspects of human object recognition and its implications for modeling recognition performance will be discussed. The second part of this chapter presents a brief summary of the current view of visual processing in the brain stemming from neurophysiological evidence. In addition, several more recent experiments are discussed, which point out the need to rethink the monolithic and inherently serial view of object recognition processing to accommodate a greater deal of connectivity and feedback. The final part ties results together from both psychophysical and physiological experiments in order to identify key concepts for the construction of the abstract recognition framework proposed in this thesis.

1.2 Cognitive psychophysics

The aim of cognitive psychophysics is to study human behavioral performance in order to come to an understanding of the processes involved in cognitive tasks such as learning and recognition of objects, navigation and orientation in environments, etc. This field of research has developed from traditional psychophysics, which investigates the functional relationship between sensory input and behavioural output using extremely well-controlled stimulus properties. Functional relationships that have been extensively studied over the last decades include sensitivity functions of basic properties of the human visual system (for example, sensitivity to contrast, line orientation, color, stereo cues, etc. - see Palmer [1999] for an excellent book on this topic). The advent of computer graphics and virtual reality techniques has allowed an extension of this methodology to more complex stimuli such as 3D (3D) objects, faces or scenes and consequently enabled a shift from investigation of low-level properties of the human perception-action cycle towards higher-level ones. Whereas psychophysics studies these properties across all perceptual modalities (for example vision, touch, proprioception, haptics, etc.), here the main focus will be on research in visual object recognition.

1.2.1 Structural versus view-based approaches

How our visual system represents familiar and unfamiliar 3D objects for the purpose of recognition is a difficult and much debated issue (for example, Biederman and Gerhardstein [1993], Tarr and Bülthoff [1995]). One of the core questions which a productive theory of recognition has to address is how much the internal model or representation depends on the viewing parameters. I will present two types of models regarding this issue:

- a structural model based on the decomposition of objects into 3D parts and the specifications of the interrelations between these parts
- a view-based model based on directly matching whole images or image parts to snapshot-like views stored in memory.

The structural framework for object representation predicts largely viewpoint-independent recognition performance based on a structural description of objects similar to 3D computer models as done for example in Biederman [1987], Biederman and Gerhardstein [1993], Marr [1982], Marr and Nishihara [1978]. This framework originates in the work of David Marr, whose influential book "Vision" [Marr, 1982] marked the beginnings of the field of computational vision. In Marr's view, vision

amounts to reconstruction, a hierarchical process that begins with local features and combines them into more and more complex structural, and thus largely view-independent, descriptions of objects (see also chapter 2). In contrast to the structural view of object recognition, the second group of recognition models is viewpoint-dependent, suggesting that objects are stored as a collection of views captured from specific viewpoints (see, for example, Bülthoff and Edelman [1992], Bülthoff et al. [1995], Rock and DiVita [1987], Tarr and Bülthoff [1998], Ullman [1979]).

For structural representations of objects a single hierarchical and compact description is sufficient to allow recognition of an object from almost any viewpoint, since any particular view of an object can be generated during the recognition process by rotation of the internally stored object representation. The structural description of objects proposed by Marr, for example, is based on hierarchical arrangements of recognition primitives - the generalized cylinders (see Figure 2.1). One major problem with this kind of object representation is, however, that it is difficult to build 3D models from 2D (2D) images on our retina, since any such 2D image is always consistent with infinitely many 3D interpretations. Biederman's recognition-by-components (RBC, Biederman [1987]) theory is motivated by the analytical and hierarchical recognition scheme developed by Marr. It represents objects with a set of 36 geometric shape-primitives called "geons". Specific (so-called non-accidental, see also Lowe [1985, 1987], Figure 1.3) properties of each geon, such as rotational symmetry or co-linearity, are invariant over a wide range of transformations such as rotation, translation or scaling, which makes them reliable diagnostic features for recognition. Unlike the 3D models of Marr and Nishihara's theory, the descriptions in Biederman's theory are not based on a faithful 3D reconstruction of the object. Rather, the description simply specifies the geons that are visible in the image and the rough 2D spatial relationships among them.

One problem with the RBC theory as well as Marr's representation using generalized cylinders is that most natural objects such as birds or fish are particularly difficult to represent with geons or generalized cylinders. In real life, people can quickly distinguish between different types of birds and fish which are nearly impossible to model thoroughly with geons to the required level of detail. Another problem is that a distinction between different objects seems to be possible only at the basic level of categorization², that is, a geon-based structural description can differentiate between a chair and a table, but not necessarily between different types of chairs. In addition, if an elongated object such as a pen is shown from one end, for example, the major axis will be foreshortened. Any structural description of the object in terms of its parts will certainly result in reduced recognition performance for these kinds of models. In all other cases, however, these theories predict similar *view-invariant* recognition performance across different views.³

A second possibility for implementing object recognition consists of a representation in which multiple descriptions of an object taken from different viewpoints are stored. Such a representation is called *view-based* (for example, Bülthoff and Edelman [1992], Poggio and Edelman [1990], Tarr and Pinker [1990], Ullman and Basri [1991]; for a more detailed account of view-based models, see Tarr and Bülthoff [1998]), and recognition in this framework is achieved by matching the retinal image to stored views in memory. View-based mechanisms for object recognition seem to require a large amount of memory compared to structural representations, and ease and accuracy of recognition will depend significantly on viewpoint familiarity. One way to overcome the storage problem is to introduce an additional matching transformation such that an image will also be recognized if it is seen from a slightly different view. In this way the amount of memory needed by the model is reduced since not every view needs to be stored. Similar matching transformations could be in place for other variations such as changes in illumination. For viewpoint changes, a variety of different mechanisms have been proposed for generalizing from familiar to unfamiliar views,

²Basic level categories - where the basic level is often also called the entry level of recognition Rosch et al. [1976] - are defined as the most abstract categories, which can be based on *visual similarity alone*. Examples of categories, which do not fulfill this requirement are "artificial objects", "means of transportation" that form more abstract "superordinate categories" often defined by *functional* instead of *visual* similarity.

³Note that a geon description also does not take into account the material or surface properties of an object and thus eliminates an important and rich source of information.

including mental rotation [Tarr and Pinker, 1990] and view interpolation [Poggio and Edelman, 1990].

Performance of view-based processing as defined here therefore critically depends on viewing parameters such as depth rotation, translation, size changes, etc. This parameter dependence has indeed been found in numerous studies ranging from recognition of novel objects to categorization of familiar objects (again, see Tarr and Bülthoff [1998]) - results, which strongly favor view-based object recognition over structural models. In the following, I will present further experimental evidence for view-based processing, which extends and supplements the view-based theory for both exemplar recognition and categorization processes along two aspects.

The first main aspect concerns recognition of highly familiar objects (faces), where one might expect different performance patterns as predicted by standard view-based theory. Such a differentiated pattern of performance could be due to the high level of visual expertise humans have for this class of objects - what one might call a "familiarity effect". Several psychophysical experiments, however, show the exact same qualitative dependency on viewing parameters for faces as for unfamiliar objects. In addition, I will discuss experiments using familiar *objects* in which participants had to identify "canonical views", that is, views which are highly informative and discriminative for a given object category. Again, performance can be best understood in the context of a view-based framework. Taken together, these two studies thus provide further support for the general applicability of view-based processes showing that view-based representations serve as the basis for recognition of both unfamiliar as well as highly familiar objects.

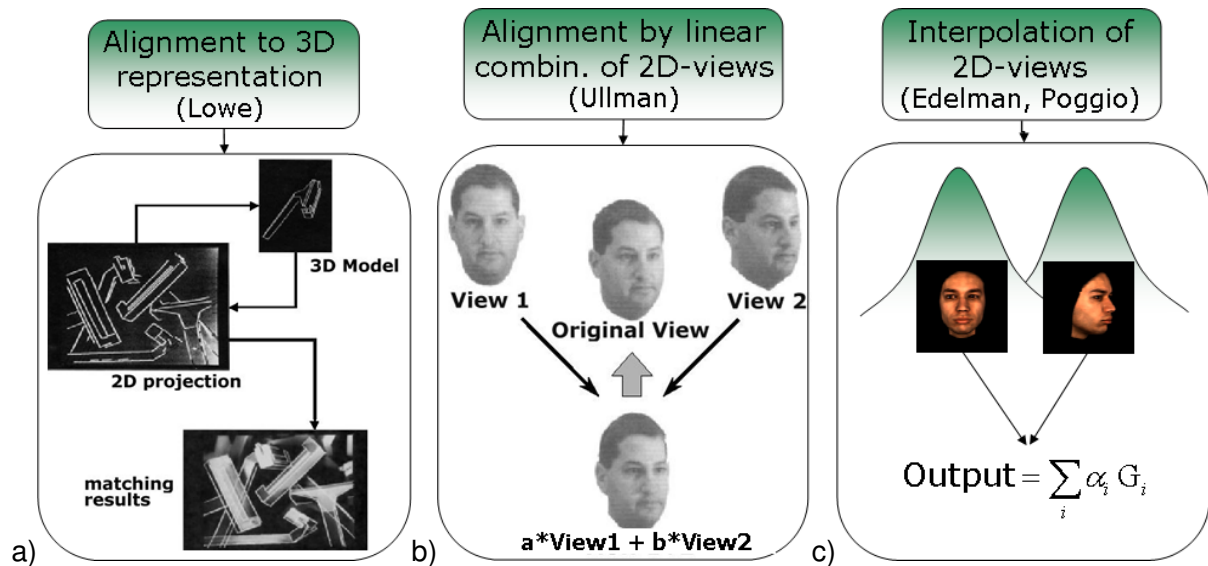
The second main aspect concerns evidence for temporal properties of object representations both during learning and recognition - a topic that only recently has begun to be explored in the context of cognitive research. I will briefly discuss two important lines of experiments that show convincingly that temporal properties are an integral part of the recognition process and in addition provide powerful cues mediating learning of visual representations. Further evidence for the importance of temporal aspects of object representations is provided by a line of *categorization* experiments, in which it is shown that both static *and* dynamic features play a role in categorical decisions.

1.2.2 View-based recognition of faces

1.2.2.1 The inter-extra-ortho experiment

One of the most important and influential psychophysical experiments, which helped to establish the view-based paradigm, is the Inter-Extra-Ortho (IEO) experiment by Bülthoff and Edelman [1992]. The study used unfamiliar objects (3D renderings of wire objects) in an old-new recognition paradigm. Participants were first trained using several views on the horizontal viewing axis and later tested on intermediate views either on the same or on a orthogonal axis (see Figure 1.4a). This design was chosen in order to enable a controlled investigation of *view generalization* for object recognition. One of the main results was that recognition performance depended critically on the viewing distance to the learned views. In addition, views between the two learned views (the inter condition) were recognized better than views which were further from the learned views on the horizontal axis (extra condition); these views could in turn be recognized much better than views on the orthogonal (ortho condition) axis - in short: inter>extra>ortho. Furthermore, the performance pattern changed in a second experiment in which training was done on the *vertical* viewing axis, suggesting that participants preferred the horizontal viewing axis over the vertical viewing axis. This was interpreted as an indicator for a *prior* of the visual system, which specifies that a horizontal viewing change should be more important than a vertical one - such a prior might be reflected in the viewing statistics of our visual world and thus might have been learned simply through exposure. The results thus demonstrated that recognition performance for unfamiliar stimuli is strongly dependent on viewpoint and provided evidence for the existence of matching priors or viewpoint preferences.

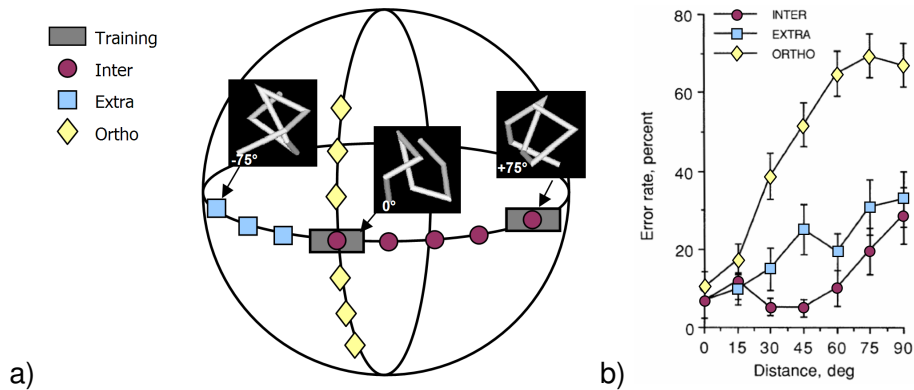
Figure 1.3: Schematic diagram explaining three different models for object recognition: a) recognition by alignment [Lowe, 1987], b) recognition by linear combination of views [Ullman and Basri, 1991] and c) recognition by interpolation of views [Poggio and Edelman, 1990, Edelman and Weinschall, 1991]



Furthermore the results of the experiment could be used to evaluate the psychophysical plausibility of several object recognition theories (see Figure 1.3), which predicted specific performance patterns for the three conditions. The first theory proposes that images of objects are aligned to 3D representations by means of extracting reliable, non-accidental features of the image and subsequent alignment of these features with the projection of a 3D model of the object [Biederman, 1987, Lowe, 1987, 1985]. As this alignment theory relies on 3D representations, consequently no difference between the three conditions should be expected if extraction of the non-accidental features is possible - in short: $\text{inter}=\text{extra}=\text{ortho}$. The second theory claims that object recognition is done by linearly combining 2D views [Ullman and Basri, 1991]. This theory of recognition by 2D alignment is based on an earlier theorem by Ullman [1979], where it was shown that any 3D object could be represented by 5 2D views together with their point correspondences. As participants in the first experiment learned two views on the horizontal axis, a recognition advantage for the horizontal axis could be expected, that is, $\text{inter}=\text{extra} > \text{ortho}$. The third theory is based on a recognition framework by Poggio and Edelman [1990] which uses *interpolation* between learned 2D views in a radial basis function network to recognize novel views. Since recognition of novel views in the inter condition would have support from *two* learned views as opposed to the extra and ortho conditions, this theory would predict $\text{inter} > \text{extra}=\text{ortho}$.

Recognition results from the first experiment in Bülthoff and Edelman [1992] are depicted in Figure 1.4b and show a differentiated performance pattern of $\text{inter} > \text{extra} > \text{ortho}$, which already rules out the first theory of alignment to 3D representations. Together with the results from the second experiment in which the learning and testing axis were swapped, it seems that the results from this series of experiments can be best understood within view-based interpolation processing. This conclusion, however, can only be reached after postulating a *horizontal view advantage* during recognition, that is, matching of views on the (viewer-centered) equator in Figure 1.4a is easier than matching views on the vertical axis. Thus, this set of experiments not only provides strong evidence that object recognition is more compatible with 2D rather than 3D object representations but in addition sheds light on the *type of processing* used during recognition.

Figure 1.4: a) Diagram showing the training and testing views for the first inter-extra-ortho experiment. b) Results from the original inter-extra-ortho experiment. Shown are three curves for the different experimental conditions (upper curve: inter, middle curve: extra, lower curve: ortho condition) (adapted from Bülthoff and Edelman [1992]).



1.2.2.2 Is face recognition view-based⁴?

In the following, I will discuss a series of experiments in which the IEO paradigm was used with one of the most familiar object classes: faces. The purpose of these experiments was to investigate whether view-based processing in addition to explaining recognition of unfamiliar objects [Bülthoff and Edelman, 1992, Tarr and Bülthoff, 1998] could also account for recognition of faces for which humans are highly trained experts.

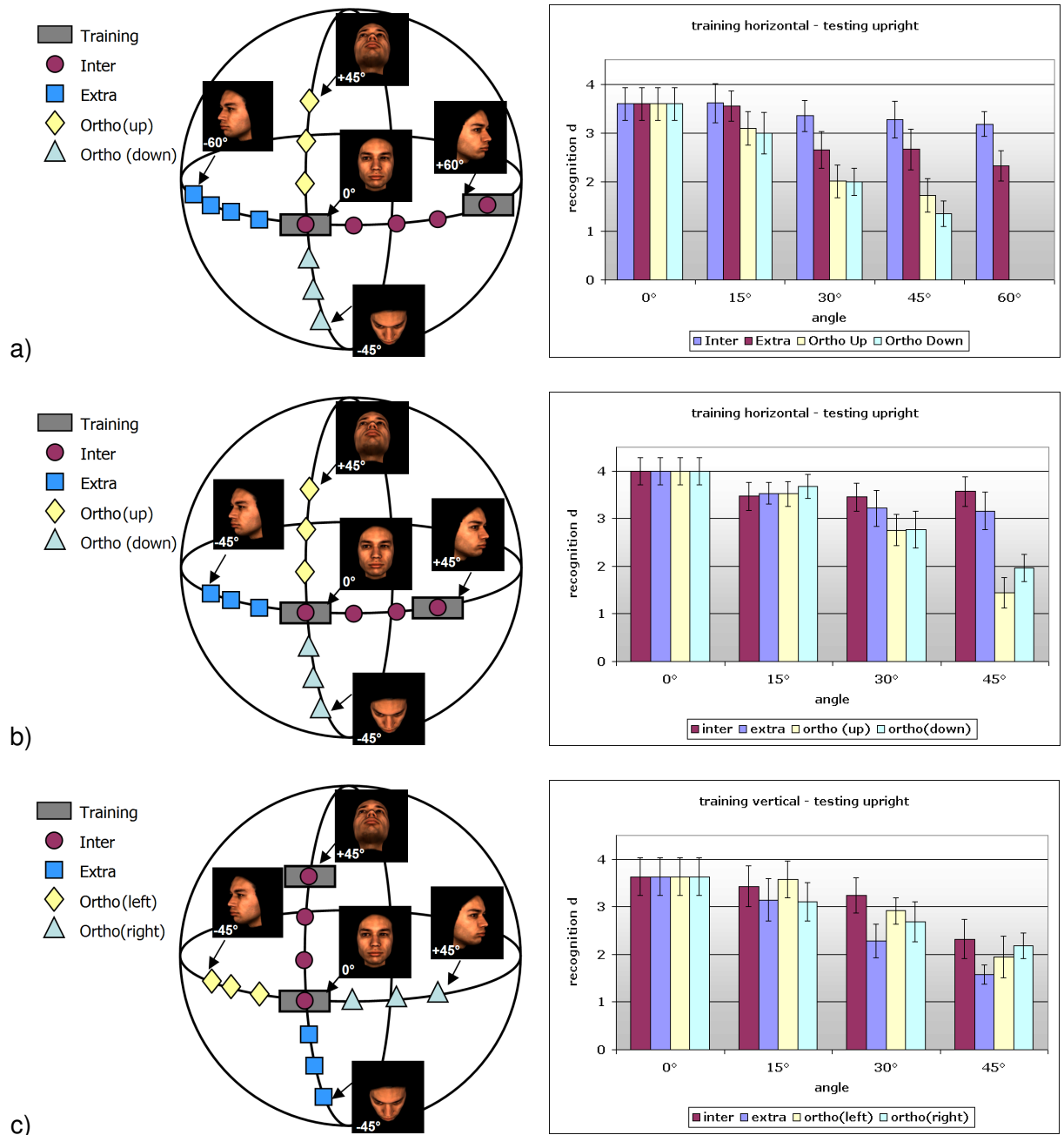
In previous studies by Troje and Bülthoff [1996], extensive experiments on the MPI face database⁵ provided evidence for viewpoint-dependence. In this study, participants had to perform a same-different task of two rendered faces, which were shown in succession on a computer screen. One important experimental manipulation consisted of the horizontal pose difference between the two images, which ranged from 0° to up to 180° (from left to right profile). Interestingly, recognition performance was not dependent on the pose of the second view but rather on the pose of the first view. This meant that *encoding* or learning of faces was dependent on the viewing angle, but recognition itself was not. Humans thus seem to be able to recognize a face seen in profile, which they had previously seen only in the frontal view. This demonstrates the impressive generalization capabilities of the human recognition system and also points to an interesting question that will also be taken up in later chapters: which features in the image would allow for the recognition of faces under 90° depth rotation?

The following study seeks to replicate the basic findings of view-based processing of Troje and Bülthoff [1996], but more importantly will go one step further by analyzing what *type* of view-based processing can support human face recognition performance. A set of psychophysical experiments were designed using the basic elements of the IEO paradigm in order to test whether recognition of such a highly familiar stimulus would rely on structural, alignment or interpolation processes. In addition, it was investigated to what degree recognition processes for faces might be affected by the extensive amount of experience and prior knowledge for this object class. The psychophysical experiments will be revisited in chapters 3 and 4, where I will present additional modeling experiments using the computational recognition system developed in this thesis.

⁴This is collaborative work with A. Schwaninger and S. Schumacher published as Wallraven et al. [2002]

⁵This database consists of highly realistic 3D laser scans of 200 individuals - the database will also be used in later recognition experiments (see also <http://faces.kyb.tuebingen.mpg.de>).

Figure 1.5: Designs and results of the first three inter-extra-ortho experiment with faces: a) standard design (large horizontal testing angle), b) standard design (symmetric design) c) exchanging the inter and ortho axis (symmetric design).



1.2.2.3 Experimental design

In total, fifty right-handed participants, aged between 19 and 40 years volunteered for the following experiments. All were undergraduate students at the University of Zürich and all reported normal or corrected-to-normal vision. Similarly to Troje and Bühlhoff [1996], the stimuli were taken from the MPI face database, from which 20 male faces were used for our experiments. Faces were rendered in a number of poses as if seen from a virtual camera placed at various points around the face while maintaining a distance of 1.35m to the center of mass. Illumination of the scene was provided by a point-light source coming from slightly above and to the right, which further enhanced depth perception of the stimuli. In addition, 20% white ambient light was added to the scene in order to avoid hard self-shadows. Each face was rendered on a black background with an image size of 512x512 pixels and further rescaled to 256x256 pixels for display purposes. The stimuli were presented on a 17" CRT screen, to which participants maintained a viewing distance of 60 cm with the help of a head rest, so that the face stimuli covered approximately 6° of the visual field.

The experiments consisted of a learning phase and a testing phase. All experiments used an old-new recognition paradigm, which meant that participants had to indicate during the testing phase whether the displayed face was one of the previously learned faces or whether it was a new face. Of the 20 faces, 10 faces were randomly selected as targets with the other 10 faces being distractors. During learning, the target faces were shown oscillating in small 1° increments both horizontally and vertically for a total of $\pm 5^\circ$ around two learning views (see Figure 1.1). Both motion sequences lasted for a total of 7.5 seconds with the order of the horizontal and vertical sequences counterbalanced across participants. The learning phase was split into four blocks (with a small break of 15 minutes between second and third block), during which the target faces were shown twice in each of the two training views. All trials in the learning phase were done without providing feedback about correct or wrong decisions.

1.2.2.4 Experiment 1a - Inter-Extra-Ortho revisited

One of the aims of the first experiment was to investigate whether the findings from the original study [Bühlhoff and Edelman, 1992] could be replicated for the stimulus class of faces. For this, learning views of 0° and 60° on the horizontal axis were chosen, following the original experimental design as closely as possible. During testing, one of the 13 views depicted in Figure 1.5a) was shown, resulting in a slightly asymmetric range of viewing angles of 60° for the horizontal axis and 45° for the vertical axis.

Results: Signal detection theory was used to determine recognition performance. The relevant measure is $d' = z(H) - z(FA)$, whereas H equals the hit rate, that is, the proportion of correctly identified targets, and FA the false alarm rate, that is, the proportion of trials in which a distractor face was incorrectly reported as a previously learned face. H and FA were converted into z-scores, that is, into standard deviation units. Individually calculated d' values were subjected to a two-factor analysis of variance (ANOVA) with condition (extra, inter, orthoUp, orthoDown) and amount of rotation (0°, 15°, 30°, 45°, 60°) as within participants factors in order to analyze the significance levels for each experimental condition. Mean d' values are shown in Figure 1.5a).

The results from the ANOVA revealed that recognition was dependent on the condition as indicated by the main effect of this factor, $F(3, 27) = 23.1$, $MSE = .354$, $p < .001$. There was also a main effect of amount of rotation, $F(3, 27) = 10.93$, $MSE = 1.500$, $p < .001$. The effect of rotation was different across conditions as indicated by the interaction between amount of rotation and condition, $F(9, 81) = 3.30$, $MSE = .462$, $p < .01$. The four conditions were compared to each other using pairwise comparisons. Recognition in the inter condition was better than in the extra condition ($p < .05$). Recognition in inter and extra conditions was better than in both ortho conditions ($p < .01$). Finally, recognition performance did not differ in the two ortho conditions ($p = .41$).

Discussion: In the following, I want to discuss whether the pattern of recognition performance

across views could be predicted by the three prominent theories of object recognition also discussed in the original study. First of all, the clear, differential effect of rotation direction that was observed seems difficult to explain by approaches using alignment of a 3D representation [Lowe, 1987, 1985]. In addition, the results question the biological plausibility of the linear combination approach [Ullman and Basri, 1991] for face recognition, because it cannot explain why performance in the inter condition was better than in the extra condition. The results can, for example, be understood by a linear interpolation within an RBF network with a horizontal view prior [Poggio and Edelman, 1990] - in chapters 3 and 4, I will present another view-based framework, which can model the results. Both of these models predict $\text{inter} > \text{extra} > \text{ortho}$, which was shown clearly in the psychophysical data. Interestingly, the pattern of recognition performance is *identical* to the original study by Bülthoff and Edelman [1992] who used paperclips and amoeboid objects in order to investigate how humans encode, represent and recognize unfamiliar objects. In contrast, this study used perhaps the most familiar object class. In this context one could conclude that familiarity with the object class does not lead to differences in viewpoint dependence and with that provides evidence for shared visual processing strategies for both familiar and unfamiliar objects.

In addition to the surprising similarity in performance between this study and that of Bülthoff and Edelman [1992] there are a few additional points that need to be discussed in particular in relation to the study by Troje and Bülthoff [1996]. First of all, there seems to be a slight decrease in recognition performance with increasing angle in the inter condition. This trend could be explained with the results of Troje and Bülthoff [1996] who found a clear effect on recognition due to the pose of the *learning* view but not of the testing view. More specifically, they found that the more extreme the angle of learning, the worse the recognition performance would be - a result, which would correspond to the trend shown in Figure 1.5a when comparing performance in the 0° and 60° view. The problem with this interpretation, however, lies in the extra condition: If there was no dependence on the *testing* view, one would not expect to find any differences between performance in the 0° condition and the views tested in the extra condition. In contrast to this, the data show a pronounced decrease in performance for the extra condition - this holds also true for the symmetric 60° view, for which this and other studies (for example, Vetter et al. [1994]) would suggest a recognition advantage.

The results thus seem difficult to reconcile with the previous results of largely viewpoint independent recognition of faces, but rather show striking similarity to the processing of *unfamiliar* objects found in Bülthoff and Edelman [1992]. In order to investigate how far this similarity in recognition performance would hold, the *vertical* training experiment used in the original study was repeated with faces.

1.2.2.5 Experiment 1b - Inter-Extra-Ortho revisited

Before presenting the results of the vertical training experiment, I will briefly discuss a second experiment, in which the findings from the previous experiment were replicated with a more *symmetric* design. In this experiment, training views in this experiment included the 0° frontal view and the 45° view on the horizontal axis, which defines angles between 0° and 45° for the inter condition, angles between -45° and 0° for the extra condition and angles on the vertical axis for the orthoUp and orthoDown condition, respectively.

Results: The mean values for d' for all participants in this experiment are shown in Figure 1.5b). First of all, recognition d' was dependent on condition type ($F(3, 27) = 22.92$, $MSE = .181$, $p < .001$). In addition, there was also a main effect of rotation angle ($F(3, 27) = 29.35$, $MSE = 0.540$, $p < .001$). The effect of rotation was different across conditions as indicated by a significant interaction effect ($F(9, 81) = 7.29$, $MSE = .324$, $p < .001$). Pairwise comparisons indicated that performance in the inter and extra conditions was better than in both ortho conditions ($p < .01$), whereas the difference in recognition performance between the inter and extra conditions was marginally significant ($p = .083$). The two ortho conditions did not differ significantly ($p = .80$).

Discussion: First of all, the results confirm the results from the previous experiment in which $\text{inter} > \text{extra} > \text{ortho}$ was found and are again most consistent with models based on view interpolation. Note, however, that in the present experiment the difference between inter and extra condition was smaller than in the previous study. This could be explained by a tendency towards *better recognition* in the extra condition. This advantage of the 45° opposite view over the 60° opposite view tested in the earlier study could be due to a benefit of the generalization performance for the 45° view. Indeed, it has been shown previously that the symmetry effect is stronger for the 45° view than for the profile view (see, for example, Hill et al. [1997]). Another possible explanation given the asymmetry of the extra and inter performance for the 45° view, consists of the smaller angle between the two learning views in this study (45°) compared to the previous study (60°). Such a smaller difference would, for example, result in a much better recognition performance in a local view interpolation framework as the influence region of the two views would overlap even more (see also Figure 1.3c). In addition to the marginally significant extra and inter conditions, a *post hoc* comparison of the 45° extra and inter views showed a significant difference. Taken together with the drop in performance for the *intermediate* views in the inter condition, this confirms our earlier results and provides further evidence for a view interpolation model. Again, these results stand in contrast to the nearly viewpoint-invariant recognition performance observed in Troje and Bühlhoff [1996]. The view-dependent interpolation hypothesis gains additional support by the clear advantage of the extra condition over the two ortho conditions which exactly follows the predicted pattern.

1.2.2.6 Experiment 2 - orthogonal training

In order to further investigate the viewpoint-dependent performance found in experiments 1a and 1b, participants in experiment 2 were trained on the *vertical* axis. Training views were therefore the 0° frontal view and the 45° view on the vertical axis (see Figure 1.5c) thus reversing training and testing axes with respect to the first experiment. Apart from the reversed axes, the experimental procedure was identical.

Results: Similar to experiment 1, both main effects of condition ($F(3, 27) = 8.49$, $\text{MSE} = .208$, $p < .001$) and rotation angle ($F(3, 27) = 25.35$, $\text{MSE} = .797$, $p < .001$) were observed. However, *no* interaction was found between the two factors ($F(9, 81) = 1.62$, $\text{MSE} = .634$, $p > .18$). Pairwise comparisons showed that recognition in the inter condition was better than in the extra condition ($p < .05$), that recognition in the two ortho conditions was better than in extra condition ($p < .05$) and finally that recognition performance did not differ between the two ortho conditions and the inter condition - in short: $\text{inter} = \text{ortho} > \text{extra}$.

Discussion: In this experiment, performance was found to be better for the vertical inter condition and both horizontal ortho conditions than for the vertical extra condition. If participants would *not* prefer horizontal views one would expect the inter condition in Experiment 2 to be better than the ortho conditions, which was not the case. Again, the experimental results are *identical* to the ones obtained in Bühlhoff and Edelman [1992], which provides strong support for viewpoint-dependent recognition based on a horizontal matching prior.

This preference for horizontal views might be caused by two factors: the first factor is based on a low-level property of the stimulus class of faces - vertical symmetry. One effect of facial symmetry is that it leads to a larger change in feature information when the face is rotated vertically than when it is rotated along the horizontal axis (see Figure 1.6a). Another visualization of this effect is shown in Figure 1.6b, which depicts the number of consistent feature trajectories that could be tracked from the frontal view to the horizontal or vertical 30° view (see chapters 3 and 4 for a more detailed description of the algorithm used). The increased amount of trajectories for the horizontal rotation show that there are more *visual changes* during a vertical rotation than during a horizontal rotation of the same angle.

The second factor that might give rise to the horizontal matching prior is based on a more

Figure 1.6: Effects of horizontal and vertical rotation on images faces. a) For a vertical rotation of 45° , many distinct features of the face are severely foreshortened or even lost as opposed to a horizontal rotation of the same angle, where the features are still visible. b) Tracking of features across 30° horizontal and vertical rotation.



cognitive perspective and deals with the viewing statistics or previous exposure to views on the viewing sphere. The distribution of views on the viewing sphere within average human experience will primarily focus on upright faces and in addition exhibit a much broader distribution on the horizontal axis than on the vertical axis, as humans as earthbound creatures usually see other humans at eye-level.⁶

The same two factors might also be the cause of the decrease in recognition performance which was observed for the 45° *training* view in the inter condition. This effect is quite pronounced especially with respect to Experiment 1, in which a much smaller difference for the two learning views was observed during testing. Again, this might be due to different amounts of discriminative visual information that are available on the horizontal versus the vertical axis. As with the horizontal view advantage during recognition, however, the possibility of a horizontal view advantage (or similarly of a vertical disadvantage) during *encoding* cannot be discounted. Such an encoding pattern would result in differences in how a representation of a visual stimulus is formed during training depending on its position on the viewing sphere and could also result from our prior visual experience with the stimulus class of faces.

In order to test which of the two factor contributes most to the pattern of human performance observed in the first two experiments, Experiments 3 and 4 were run with the same design as Experiments 1 and 2 but with *inverted* faces. This simple manipulation has been shown to severely affect by recognition and processing of faces (Yin [1969]; see also later experiments in this section on configuration and components as well as Rock [1973], Thompson [1980]). Presentation of inverted stimuli during testing therefore allows to assess whether any effects observed in the previous two experiments were due to *class-specific* learning effects or due to simpler *low-level properties* of facial images such as vertical symmetry.

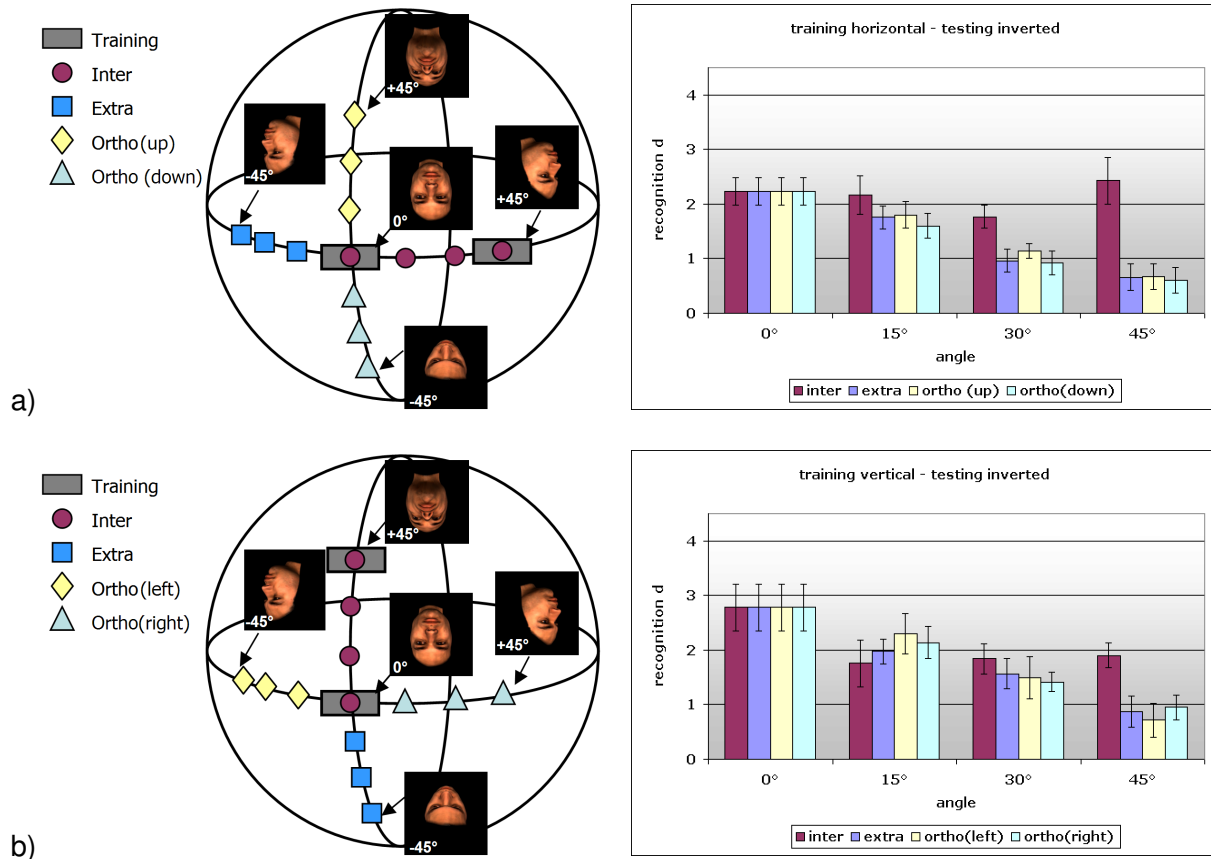
1.2.2.7 Experiments 3 and 4 - inverted faces

The experimental design was the same as in experiments 1 and 2, but with inverted faces as stimuli in the testing phase (see Figure 1.7a+b).

Results: Similar to the previous experiments, both a main effect of condition ($F(3, 27) = 13.00$, $MSE = .434$, $p < .001$) and rotation angle ($F(3, 27) = 15.07$, $MSE = 0.769$, $p < .001$) were found with a significant interaction between the two ($F(9,81) = 4.28$, $MSE = .948$, $p < .05$). In pairwise comparisons, recognition in the inter condition was better than in the extra condition and the two ortho conditions (all $p < .01$). However, recognition performance did not differ between the two ortho conditions and the extra condition.

⁶This hypothesis can be seen as a concretization of the comment in the original study by Bülthoff and Edelman [1992] where the authors remark "that the bias in favor of the horizontal plane is ecologically justified, since it is probably more useful to generalize recognition to a side view than to the top or bottom views."

Figure 1.7: Designs and results of the two inter-extra-ortho experiment with inverted faces: a) same as 1.5b) but with inverted faces, b) same as 1.5c) but with inverted faces.



In experiment 4, no main effect of condition ($F(3, 27) = 1.76$, $MSE = .376$, $p = .18$) was found but a significant main effect of rotation angle ($F(3, 27) = 15.43$, $MSE = 1.314$, $p < .001$) with a significant interaction ($F(9,81) = 2.47$, $MSE = .418$, $p < .05$). Pairwise comparisons indicated only one significant effect of recognition in which the inter condition was better than the ortho(left) recognition performance ($p < .05$).

Discussion: As a first result, performance in both experiments was significantly lower than in the upright conditions of the previous experiments, which illustrates the effects of inversion on face processing. If the preference for horizontal views was purely based on low-level symmetry, one would expect no change in the relations between the inter, ortho and extra conditions. The results from experiment 3, however, show a significant change in the performance pattern with both ortho conditions equal to the extra condition. Thus, a simple inversion seems sufficient to destroy the horizontal matching prior. This indicates that this prior is largely dependent on visual experience with the stimulus class of faces rather than on low-level information which was preserved in this manipulation.

Although the data is not consistently interpretable, experiment 4 seems to support this hypothesis of a learned matching prior, which is suggested by the trend of better recognition performance of the inter condition versus the ortho conditions. One possibility for the inconclusive results could be a small but noticeable residual effect of the low-level preference for horizontal matches, which interacts with the inversion effect. Note also, that recognition of the learned 45° inter view is significantly lower than that of the learned 0° view - an effect that was observed also in experiment 2 and that could be due either to reduced visual information or to differential encoding fidelity for the

horizontal versus the vertical axis.

1.2.2.8 General discussion

In summary, all of the above results are difficult to explain with approaches using alignment of a 3D representation [Lowe, 1987, 1985], as a differential effect of rotation direction on recognition performance would not be expected - even more so, since the actual range of rotation angles used in the experiments is comparatively small. Furthermore, it is difficult to envisage such a strong effect of learning axis (experiment 1 versus experiment 2) on recognition performance - especially with highly over-learned stimuli as faces, for which it can be assumed that participants had the chance to form a sufficiently detailed 3D representation of the faces during training. A third problem for modeling with view-invariant representations is given by the drastic performance differences between the inverted and upright conditions in both the horizontal and vertical learning condition. If faces are treated as an invariant 3D representation, turning them upside down should have no noticeable effect on recognition performance⁷.

Since the results of the 5 experiments speak strongly in favor of viewpoint-dependent processing during recognition, one has to ask what the critical difference is between this study and the one of Troje and Bühlhoff [1996] which only found viewpoint-dependency for learning or encoding of views. The most plausible explanation lies in the different tasks of the two studies: Troje and Bühlhoff [1996] used a more rapid same-different recognition task resulting in a much decreased memory load compared to the old-new paradigm of this study. This in turn might lead the visual system to use a different processing strategy. In addition, the IEO experiments used *two* training views for each face, which could give rise to a different, potentially richer representation of each identity. It would be interesting to incorporate *different* learning and testing views into the inter-extra-ortho paradigm to more explicitly decouple the influence of encoding and recall on recognition performance. In addition, it would also be worthwhile to extend the study by Troje and Bühlhoff [1996] to also include the vertical viewing axis as well as to use longer inter-stimulus intervals in order to investigate the influence of memory.

In summary, the experiments conducted in this study show that recognition of faces in an old-new paradigms can be explained surprisingly well by view-based interpolation approaches [Poggio and Edelman, 1990], where the pattern of recognition results depends heavily on experience and familiarity with the stimulus class. It could thus be demonstrated that view-based recognition has a significant effect not only for unfamiliar stimuli but also for a class of highly familiar stimuli - a stimulus class, for which humans are experts.

1.2.3 The canonical view

As one of the earliest examples of view-dependent object recognition performance, Palmer et al. [1981] reported that certain "canonical" views are recognized more quickly than others. These experiments were done with highly familiar objects such as coffee mugs. In a recent study, Blanz et al. [1999] investigated which kinds of visual attributes of objects give rise to a canonical view. The psychophysical experiments were done using interactive computer graphics to allow exploration and manipulation of objects by participants, resulting in a more natural interaction with objects than in the study by Palmer et al. [1981].

⁷One might argue that for a given object class, rarely experienced geometric transformations (viewing faces upside down) or learning effects such as introduced by the horizontal and vertical learning conditions could give rise to differential effects in recognition *time*. Nevertheless, it does not seem possible to predict the distinct pattern of recognition *errors* of human performance with the standard 3D approaches.

1.2.3.1 Experimental design

In the first of two experiments, participants were instructed to rotate an object on the computer screen such that the view was maximally informative (for example, for illustrating a brochure advertising this particular object). In the second experiment, participants mentally imagined each object based on its name and afterwards rotated the object to the mentally imagined viewpoint, which is a task that taps into long-term object representations in memory. Stimuli consisted of 3D textured objects comprising both artificial and natural common objects.

1.2.3.2 Results

One of the most important results from the first experiment was that participants selected consistent viewpoints for all objects - something which could not be explained by a uniform distribution of views (see Figure 1.8) and which confirmed the idea of canonical views for these objects. Similar to the results found in Palmer's study, participants preferred off-axis views to planar views thereby maximizing the amount of visible surfaces. Furthermore, accidental views in which important features of the objects would be fore-shortened or invisible were avoided and participants tended to select views in which all parts of the object were visible thus maximizing visual discriminability of objects.⁸ Interestingly, symmetric objects resulted in symmetric canonical views, and objects with a clear upright such as cars resulted in canonical views in the upper viewing hemisphere. For a 3D head-model, the preferred view was 32° displaced on either side of the frontal view, which was consistent with earlier studies. It is interesting to note the difference to the previous inter-extra-ortho experiment (as well as to the study by Troje and Bühlhoff [1996]) in which symmetry could *not* fully account for the observed viewpoint dependency during recognition - again, this might be due to the lower memory load required to solve this canonical view task as the objects were always visible on the screen. Finally, participants tended not to choose views which were optimal for grasping but instead chose views which were *visually* informative.

In the second experiment, participants first imagined an object and then rotated it on the computer screen to the imagined viewpoint - a task which taps into the long-term representation of objects using semantic memory access. Again, a uniform distribution of views could not explain the observed view preferences. The selected views, however, differed between the two experiments. More specifically, participants selected more frontal and side views than in the first experiment, which also meant that accidental views were not longer avoided. Finally, selected viewpoints were asymmetrically distributed on the viewing sphere indicating the influence of the task on the mental representation of the object⁹.

1.2.3.3 Discussion

In both experiments, there was a clear tendency of participants to prefer a specific viewpoint, which confirms the canonical view effect. The different results of the two experiments might be due to a trade-off between diagnosticity and simplicity of views. In the first task, participants could use the full visual information about the objects, whereas in the second task the image originated in long-term memory, which might have a limited amount of storage for views thus potentially limiting view complexity. Nevertheless, the difference between the two experiments shows a clear influence of the task on the chosen canonical views.

The general result from both experiments is that object familiarity is weighted more than functional or geometric considerations. Such a view preference according to the statistical properties

⁸The experiment also included a few unfamiliar objects (such as "geons" and "paperclips"), that seemed not to have distinct canonical views although the participants' choice of viewpoints was still restricted by the strategy to avoid accidental views and favoring maximum visibility of parts.

⁹Unfortunately, the second experiment did not include the 3D head-model, which could have clarified the role of memory recall mentioned above.

Figure 1.8: Example of canonical views for several objects (adapted from Blanz et al. [1999]).

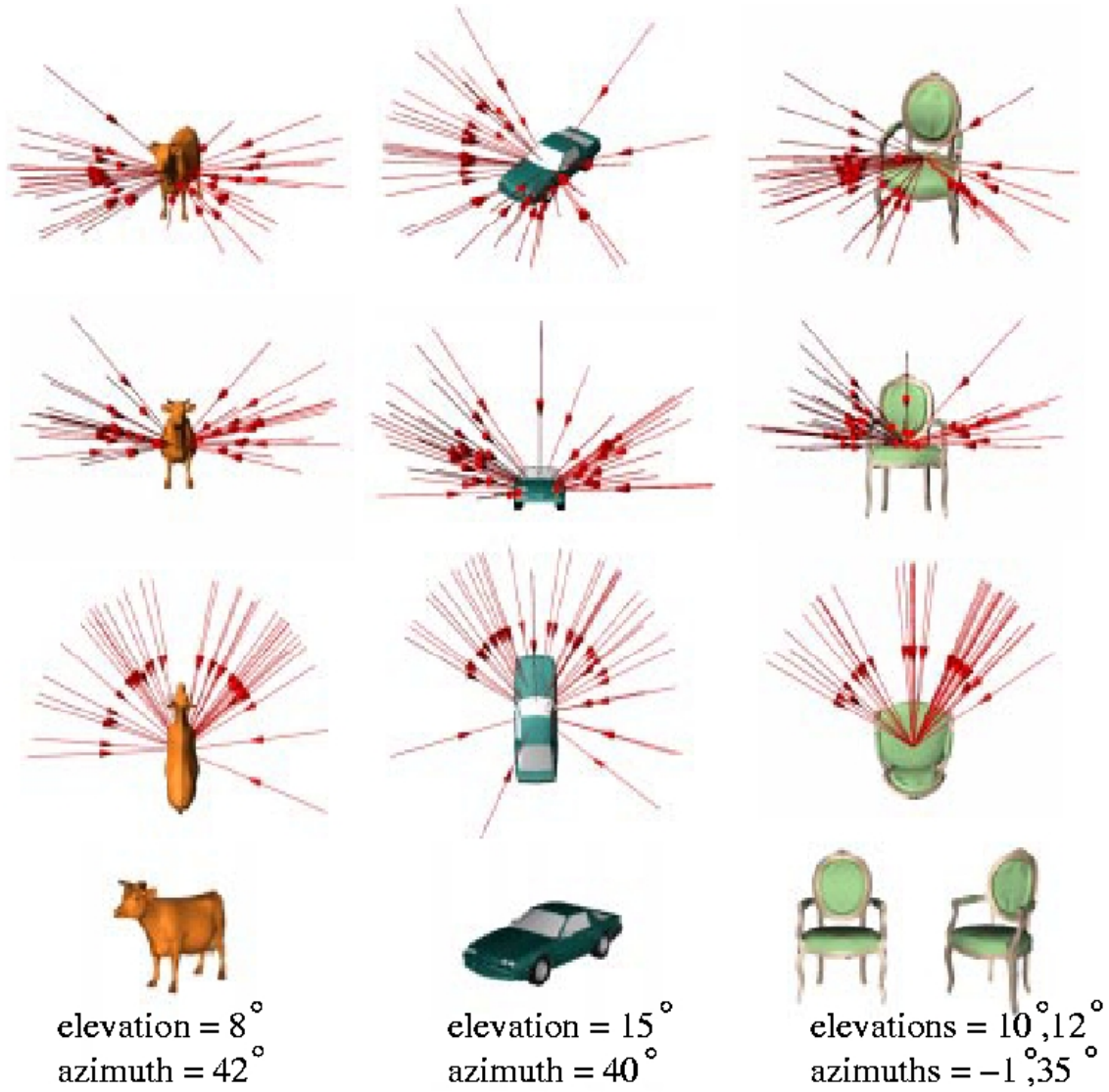
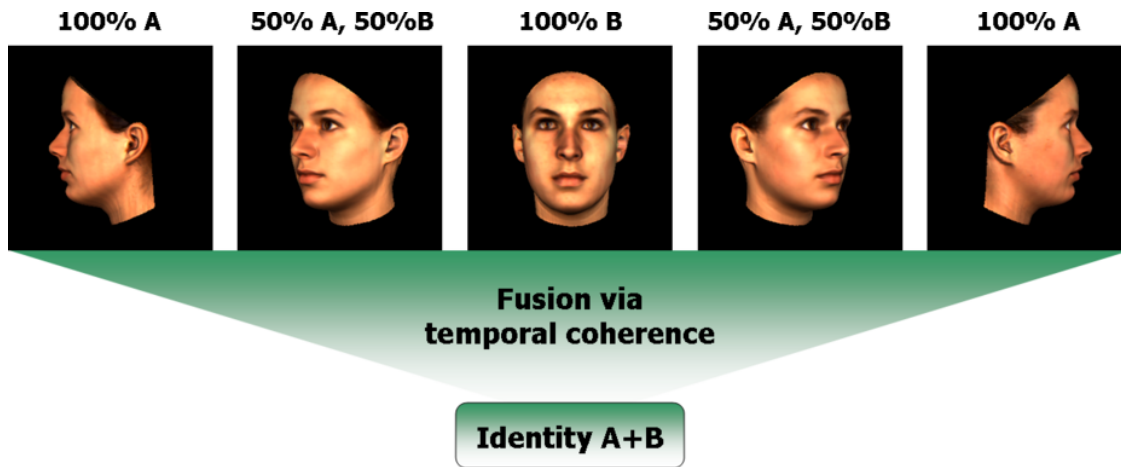


Figure 1.9: Example morph sequence from experiment on temporal contiguity. A face undergoes morphing during rotation.



over repeated exposures is precisely the result that a view-based theory of object recognition - such as discussed in the previous experiment - would predict. Interestingly, the results were not influenced by the grasping affordance of the objects but they rather seemed based on visual properties alone. This observation might be the result of the purely visual task - it would be interesting to continue the same line of experiments to investigate whether there might be haptic or visuo-haptic canonical views (see also the outlook in chapter 8).

1.2.4 Temporal aspects of object learning

Since the world around us is constantly changing and we are moving around in the world, the visual input received by the brain is inherently dynamic. An important question following from this is whether the temporal quality of the visual input influences our recognition capabilities. This issue becomes especially important in the context of object learning in a view-based framework, where the visual system faces the problem of how to link different views of one object to create a consistent and coherent object entity, especially when these views are very different from each other.

One solution to this problem is the observation that in real life we seldom see only isolated snapshots of objects. Usually novel objects are explored either actively through manipulation by our hands or by walking around them. This way a sequence of images is acquired that is gradually changing from the initial view of the object to a very different one within a short period of time (temporal contiguity). This general observation motivates the following question: Does the human visual system use temporal contiguity to build mental representation of objects, thereby linking potentially very different views together?

Wallis and Bühlhoff [2001] have proposed that our visual system uses temporal contiguity to associate images with one particular object. This temporal association hypothesis was investigated in two experiments [Wallis and Bühlhoff, 2001, Wallis, 2002] by asking whether participants might be induced to confuse two different faces if these faces were associated in a coherent temporal sequence.

1.2.4.1 Experimental design - morphs

Twelve faces from the MPI face database were used as stimuli (Figure 1.9). Using a technique by Blanz et al. [1999], 3D morphs between all possible combinations of face pairs within each

set were created. A sequence of five images was produced for each pair of faces A and B. A forward sequence consisted of a left profile view (-90°) of the original face A, a -45° view of morph A→B (the average of face A and B), a frontal view (0°) of face B, a $+45^\circ$ view of morph A→B, and finally a right profile ($+90^\circ$) of face A (Figure 1.9). A backward sequence showed the same images in reversed order. The training sequence consisted of the forward sequence followed by a backward sequence, followed by a forward sequence and a backward sequence again. Thus, in one training sequence the participants saw a face turning from left to right and back again twice. The complementary (swapping identities A and B) training sequence was also created, resulting in a total of 12 (all possible pairings of each group twice) training sequences per training set.

The participants were divided into two groups. Each group underwent a different kind of training with the stimuli before testing. In training with sequential presentation each participant in the first group saw all 36 (3 training sets · 12 sequences) stimulus sequences presented in random order. Each image in the sequence was shown for 300 ms, followed immediately by the next to mimic a face rotating back and forth. Training with simultaneous presentation consisted of all the faces and morphs from one forward sequence shown *together* on the computer screen for the total time of 6000 ms. After training all the participants did a delayed match-to-sample task in which, after a fixation period, one image was shown followed by a mask and then a second image followed by a mask again. The participants had to decide whether the two images were different views of the same face or not. Half of the 384 trials presented matches whereas in the other 192 trials, 96 of the test face pairs belonged to the same training set (within set, WS) and 96 to different training sets (between set, BS).

Test images consisted always of a frontal view followed by a profile view or vice versa (see the previous discussion on the study by Troje and Bühlhoff [1996]). All participants completed four blocks, two on the first day and two on the following day with each block consisting of a training phase followed by a testing phase.

1.2.4.2 Results+Discussion - morphs

If views of objects are associated based on temporal contiguity, then training with a sequential presentation should cause the images grouped in one training sequence to be fused together as views of a single object. After such training, participants in the testing phase would be expected to fail in the same-different task for faces that were linked together in a training sequence more often than for faces that were not. Training with simultaneous presentation was included to rule out the possibility that the morphs alone were sufficient for the training effect in which case an effect should appear after both training procedures.

Indeed, the results of the experiment confirmed that participants were more likely to confuse those faces that had been associated temporally in a sequence (WS). An analysis of the results indicated significantly poorer discrimination performance for faces learned in the WS condition than for faces learned in the BS condition. Performance for BS faces appeared to increase slightly over the four blocks due to increased familiarity with the stimuli, this was not the case, however, for WS faces. Thus, the difference in performance between BS and for WS comparisons was more pronounced on the second day. The results after training with sequential presentation therefore support the predictions of the temporal association hypothesis and show how participants learned to associate views of different faces with each other *without* any explicit training .

The results from the second group of participants, who had been trained with simultaneous presentations of the images, indicated that simultaneous presentation of views of different faces alone was *not* sufficient to cause these faces to be associated with each other. This is interesting because it demonstrates that temporal processing is operating under different constraints during *sequential* presentation than for example during saccadic eye movements between the images (as necessary in the side-by-side presentation), which would also form a potential temporal window in which association could take place. This agrees with the general assumption that overt saccadic

movements trigger new analysis cycles and result in diverting attentional resources to a different location of a scene or to a different object¹⁰.

In spite of the evidence presented, there is, however, one potential problem with this study, which is given by the morphing procedure used to create the trainings sequences for the first group. The results might be simply due to the *morphing* itself and not to any temporal association effect. In addition, the question remains whether the temporal association effects are due to a purely temporal process - as indicated by the term temporal *contiguity* - or whether the association process might also have a similarity (a "spatial") component. In the latter case one should expect that dissimilar images in a sequence would *reduce* temporal association. Given the results from the simultaneous presentation experiment, however, the possibility that the effect is *purely* based on such a similarity effect is unlikely, as association by similarity alone would have resulted in a similar effect as in the first sequential presentation experiment. In order to investigate the role of morphing and the nature of the association effect, a second set of experiments was conducted in a follow-up study [Wallis, 2002]:

1.2.4.3 Experimental design - different faces

The design of the first experiment closely followed the previous study for the morph group. The major difference was that the morphed learning sequences were replaced with sequences of *different* faces rotating. First, three groups of 5 faces were created from the MPI face database, where each face again was available from left to right profile in 45° steps. For training, rotation sequences were assembled from each group by taking each pose of the rotation from a *different* face within the group such that some form of temporal *continuity* was retained. The task was the same as in the previous study in which participants had to decide whether two face images differing by 90° belonged to the same face or not. Half of the test trials were matches, whereas the other half was split between WS face images (faces from the same group) and BS face images (faces from different groups). Participants completed three blocks of training and testing with a 10 minute break between blocks.

In the second experiment, training sequences were created by *scrambling* the poses in the sequence such that at most two consecutive images showed a consistent and smooth rotation (of 45°). The remaining experimental parameters were the same as for the first experiment. The purpose of this experiment was to test whether temporal association based on temporal *contiguity* could still be detected even when the spatial similarity between consecutive images was low.

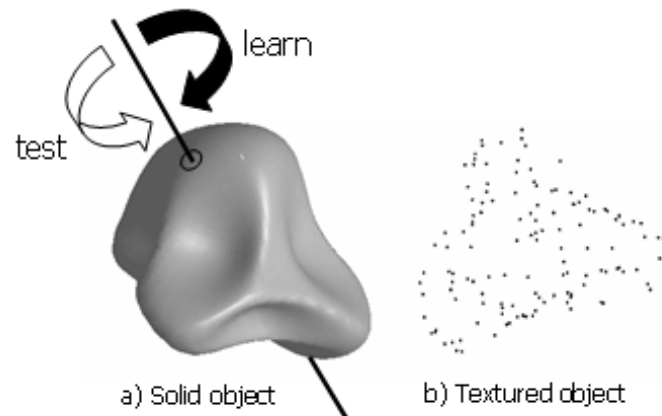
1.2.4.4 Results + Discussion - different faces

The main result from the first experiment was that WS scores were significantly lower than BS scores indicating that temporal association, indeed, is possible even with more dissimilar sequences. However, the relative effects of temporal association on the two test conditions were *much reduced* when compared to the morphed sequences in the first study. This is already a first indication that association might be influenced by spatial similarity as well as temporal contiguity.

In the second experiment, no significant main effects were found for a scrambled presentation of images, which destroyed a consistent rotation interpretation and made consecutive images spatially dissimilar but which otherwise should have left temporal *contiguity* intact. However, over the course of three blocks, a significant trend towards a slow dissociation between the two test conditions could be detected. The authors interpreted this as a sign that temporal association can take place under such conditions - albeit at a much slower rate (see also Miyashita [1988] for a similar long-term learning study with monkeys).

¹⁰This phenomenon is studied under the name of trans-saccadic integration.

Figure 1.10: Stimuli used in experiments by Stone [1998, 1999]. Shown are two renderings of the same object a) shaded, b) textured, which was learned in a rotation sequence around an arbitrary axis. Recognition performance was affected when the sequences had to be recognized for the reverse rotation direction.



1.2.4.5 General discussion

Summarizing the two studies, one can conclude that learning of object representations is strongly influenced by the temporal properties of the visual input. One successful strategy, how the brain might solve the task of building consistent object representations - even under large changes in viewing condition - seems to be to assign consecutive images to the same object. Such a strategy is not only influenced by temporal parameters but also to a significant degree by similarity properties of the input - arbitrary images seem to be much harder to learn suggesting a significant influence of the *spatial* component of visual input. I would thus interpret these results by Wallis et al. not so much as a sign of temporal *contiguity* but rather to be consistent with the extended concept of *spatio-temporal continuity* - integration of images that are below a certain similarity threshold and are presented within a certain time window.

Another question one might ask is: do these results generalize from faces to other objects? After all, faces are special objects [Kanwisher et al., 1997] for which humans are experts. In this context, the work done by Sinha and Poggio [1996], who investigated the role of learning in form perception using wire-like objects, showed that participants could build up different mental representations of the shape of ambiguous objects, depending on which views were temporally associated during the training phase. Thus, the temporal association effect does not seem to be restricted to the special category of faces (see also [Stone, 1998, 1999, Wallis, 2002] and the next section for additional discussion on this topic).

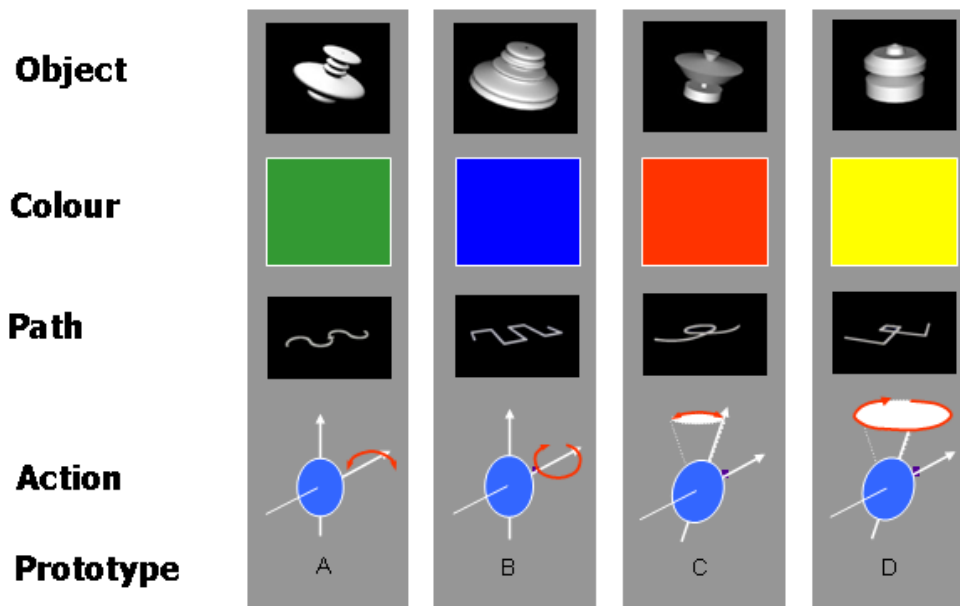
1.2.5 Temporal aspects of object recognition

In addition to this experiment showing how temporal association is used during acquisition of object models, there is also evidence that *temporal characteristics* can become part of object representations.

1.2.5.1 Object representations are spatio-temporal

In studies by Stone [1998, 1999], four different amorphous 3D objects were used as training and testing objects. These objects were shown in a sequence while rotating in a tumbling motion either clockwise or counterclockwise around an arbitrary axis. Participants first had to learn the target

Figure 1.11: Prototype stimuli as used in the categorization experiment. Each prototype is defined by four feature dimensions: two static (shape and color) and two dynamic (path and action).



objects rotating in one particular direction. After learning, half of the objects reversed rotation direction - that is, the sequence was played backwards (see Figure 1.10). For these sequences large performance drops both with increased response times and error rates were observed. Furthermore, these results could be obtained not only with gray-scale images but also with point-light renderings of the objects [Stone, 1999]. From a purely computational point of view, both directions of rotations specify the same set of views and in addition also determine the exact same 3D structure in a structure-from-motion framework. The results thus show that object representations are inherently *spatio-temporal*, which goes beyond the traditional assumption of static representations for rigid objects.

In a recent study by Vuong and Tarr [2004], participants had to recognize views of both "easy" and "hard" novel objects in a *short-term* memory task. In each trial, participants were first shown a sequence of a rotating object (rotating in one of two directions) which was followed by a static test image. This test image could either depict the same or a different object. In addition, the pose of the object could either be taken from poses shown in the sequence or from a preceding or succeeding pose, where the angular differences between all tested poses was the same. The results showed a clear effect of temporal information in that poses towards the end of the sequence were recognized faster and more accurately. In addition, preceding poses showed poorer performance than succeeding poses. Further experiments showed that this differential effect of rotation direction could not be explained by viewpoint-dependent recognition for the objects and that randomized, sequential presentation of the images destroyed the effect. The results of this study demonstrate that recognition in a spatio-temporal context is qualitatively different from a purely static context.

Taken together, these two studies provide strong evidence that the human visual system makes use of temporal properties of the visual input both for short-term as well as long-term recognition of objects - a fact that previous theories of object recognition have so far largely neglected.

1.2.5.2 Object categorization relies on spatial and temporal cues¹¹

Whereas in the previous studies the focus was on recognition, in the following set of experiments it will be investigated to what degree spatio-temporal cues play a role in forming perceptual categories and for *categorizing* new objects. Object recognition and object categorization have been typically studied either with static objects (that is, with no dynamic information) or with dynamically presented point-light displays (that is, with little shape information). But how are dynamic and static features integrated in object representation? In particular, what is the role of movement for object recognition and categorization when the object is presented in its entirety, that is, when static and dynamic features are presented together? Here, I want to briefly discuss three experiments where novel objects were categorized on the basis of two spatial properties (color and shape), and two dynamic properties (action and path). The "action" of an object referred to its intrinsic motion pattern, whereas "path" referred to an object's extrinsic motion pattern, that is, the route an object took with respect to an external reference frame. The task for the participant was first to learn to categorize prototype objects and second, to categorize novel exemplar objects, which varied in number and type of properties which they shared with the prototype (see Figure 1.11 for examples of prototypes and object properties).

1.2.5.3 Experiment 1 - categorization of objects

Experiment 1 consisted of two phases: a learning phase with feedback followed by a test phase without feedback. In the learning phase, each subject was shown a movie and instructed to learn the object and press one of two buttons indicating one of the two learned prototypes. Six trials were presented and participants received feedback for each trial. Test stimuli were derived from the prototype either by changing a dynamic feature or a static feature with feature changes counterbalanced across all participants. In the test phase, the task for the participant was to correctly associate each exemplar under either one or two feature changes with the appropriate category (a two-alternative forced-choice paradigm). The error rates across participants for all trials were then calculated as a *bias* from the actual percentage difference between the exemplar and the prototype. The mean percentage bias for each feature change is presented in Figure 1.12a. A positive bias means that the participants were sensitive to this feature and tended to over-estimate changes from the prototype with that feature, whereas a negative bias shows that the participants used changes to this feature in their categorization decisions.

A two-way ANOVA using one between factor (paired prototypes learned) and one within factor (feature changes from prototype) was conducted on the percent bias responses across all trials (that is, learning and test trials). No effect of paired prototype learned was found ($F(1,38) = 1.49$, n.s.). A main effect of feature was found ($F(4,152) = 2.96$, $p < 0.05$), for which a *post hoc* test revealed that the "path" feature was significantly different from all other features ($p < 0.05$) except for the combined shape-color/action-path (SC_AP) feature. There were no other differences between the features. In addition, no interaction between the factors ($F(4,152) = 0.62$, n.s.) was found. Using a non-parametric sign test, the extent of the response bias to each feature change were compared against no bias (essentially against a bias of zero meaning equal weighting for each feature). None of the shape, colour and action features showed any significant difference from zero indicating no evidence of a bias to these features (shape, $Z = 1.11$; colour, $Z = 0.64$; action, $Z = 0.81$). Also the exchange of two static or two dynamic features did not result in any bias (SC_AP, $Z = 1.31$). On the other hand, responses to the path feature were significantly different from zero, indicating a response bias to this feature ($Z = 3.95$, $p = 0.0001$). In a further analysis, the bias to each feature change against no bias was compared only for the responses to the non-feedback *test trials*. This was done to ensure that the findings were not due to the feedback given in the learning phase of the experiment. Using the sign test no evidence of a bias for any of the feature

¹¹This is collaborative work with S. Huber and F. Newell published as Wallraven et al. [2001], Newell et al. [2004]

changes was found, except for the "path" feature which was, again, significantly different from zero ($Z=2.60$, $p=.009$).

The two major findings from experiment 1 were thus that "path" seemed to be ignored for categorization and that participants were as likely to use dynamic cues as they were to use static cues for object categorization. One reason for the negative "path" bias might be that the path and object takes is *per se* not as indicative of category membership as the other features (we seldom would classify for example, an animal as a predator by using its walking path). The following two experiments investigated further whether path was at all an accessible feature available for encoding and whether it might have been overshadowed by the stronger action feature.

1.2.5.4 Experiment 2 - similarity judgments

In the second experiment, differences in perceptual saliency of the four features were tested in order to ensure that no single feature (in particular "path") would be overpowered by the other features. In any one trial, participants had to rate the similarity of two objects using a scale from 1 to 7, where a rating of 1 indicated a high degree of similarity. One of the objects was always a prototype object and the other an exemplar object. The exemplar differed from the prototype either in 1 feature or 3 features. The two objects were presented next to each other and started moving at the same time. Participants had to participate in four blocks which differed in the pair of prototypes used (AB, AC, BD, or CD). In each block participants conducted 2 similarity ratings for each 1 and 3 feature change and each prototype resulting in 32 trials per block. In particular, if all features would have similar perceptual saliency, the similarity ratings for both a single feature change and a three feature change should show *no* difference between feature types.

The mean ratings per feature for 12 participants are shown in Figure 1.12b. There was a significant difference between the one and three feature changes ($F(3,33) = 3.89$, $p < 0.05$), whereas the only significant effect for one feature changes was that color changes were rated as more similar than action changes ($p < 0.05$). There were no significant differences between the three feature changes, however. The main result of the second experiment with respect to the previous experiment was that "path" was a feature that was explicitly used during this similarity task, which showed that it seems available for perceptual encoding. In total, there seems to be a largely uniform saliency for all four features, which again confirms our previous result showing that both static and dynamic cues play a role during cognitive tasks.¹²

1.2.5.5 Experiment 3 - categorization experiment with redundant action feature

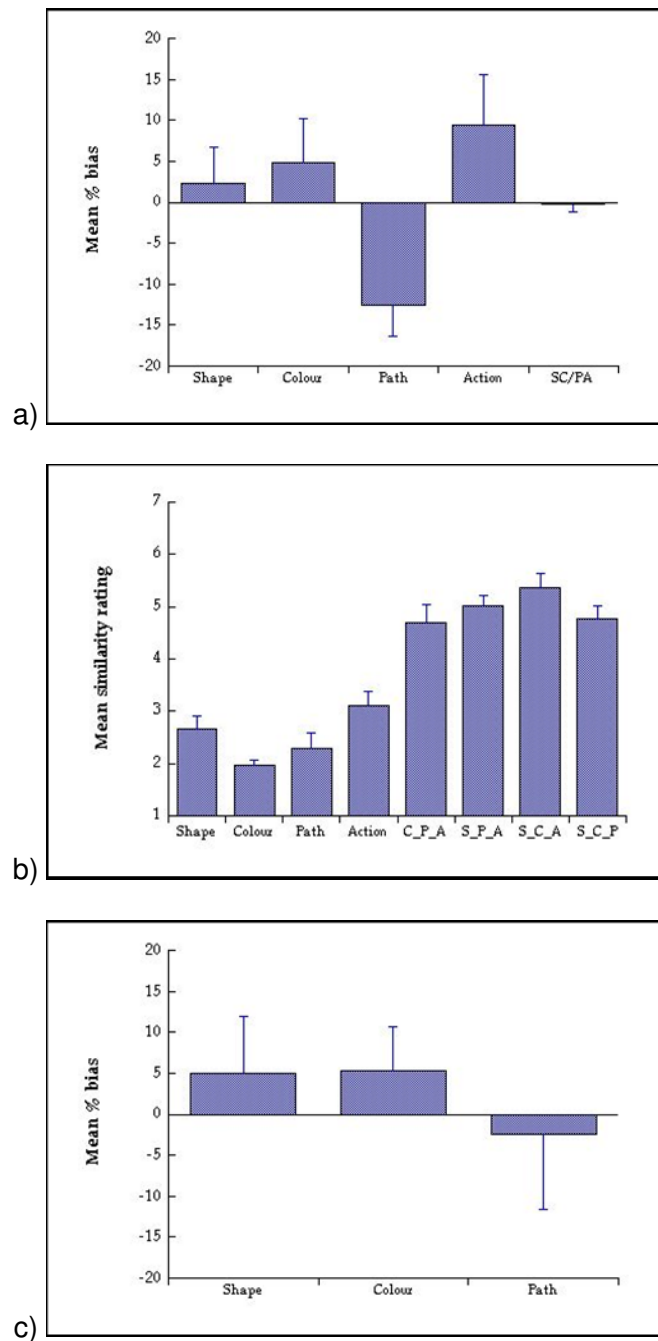
To test the hypothesis, whether path might be overshadowed by a stronger action feature, a third experiment was conducted in which the displayed action was the same for all prototypes. This reduced the set of discriminating features to "color", "shape" and "path". The design followed that outlined in experiment 1 for the AC, BD group of participants only.

The mean percentage bias for 16 participants calculated as in experiment 1 for each feature change is presented in Figure 1.12c. A one-way, repeated measures ANOVA was conducted on the bias scores using feature type as the factor. There were 3 levels to the feature type indicating the types of feature differences between the exemplar and prototype ("shape", "color", and "path"). No effect of feature type ($F(1,3) < 1$) was found. In addition, no evidence of a bias for any of the features ("shape", $Z=0.25$, n.s.; "color", $Z=1.25$, n.s.; "path", $Z=.25$, n.s.) could be detected.¹³

¹²The difference for the color change seems not to have an overall effect as it is not present in the three feature changes.

¹³For the separate non-feedback trials we found a main effect of feature type ($F(1,3)=2.99$, $p<0.05$). A post-hoc test revealed that the biases for path and colour were significantly different ($p<0.05$) and sign tests showed that responses to any of the features were not biased from zero (shape, $Z=-0.25$, n.s.; colour, $Z=1.75$, n.s. and path, $Z=0.25$, n.s.). This effect might be due to our instructions and the disproportionately high weighting of static cues in this experiment.

Figure 1.12: Results from categorization experiments presented as a function of feature changes. a) percentage of bias, b) similarity ratings, c) percentage of bias with one redundant dynamic feature



For this (easier) task, no evidence of a bias for any of the features was found showing that "path" was used as readily as shape and color properties for categorization of objects. Thus, categorization is possible using "path" as a discriminative feature. Again, there was no predominance of static cues over dynamic cues.

1.2.5.6 General discussion

Traditionally, categorization and recognition have often been studied exclusively in the static or dynamic domain. In these experiments on categorization both types of cues were brought together in a perceptually relevant and realistic task of categorization of novel objects. The main result of this study was that when presented with a number of static and dynamic cues, participants readily made use of both types of information. First of all, this shows that all of these features are accessible to visual memory and thus are part of the stored and learned representations of these objects. This again supports the results from Wallis and Bülthoff [1999], Wallis [2002], where it was shown that the dynamic properties of objects were encoded during learning. It also supports the view that without any prior information about cue saliency there is no intrinsic advantage of static cues in such a task. The results, however, are in contrast to the findings of Mak and Vera [1999], who reported that for older children, motion information was important in tasks with shapes that are commonly seen as moving in the real-world, but *not* in static geometric shapes. For younger children on the other hand, motion information was used for both shape types. Consequently, it can be suggested that it is only through the use of *novel* objects that the exact role of motion for general object recognition can be determined.

The second main finding was that there seems to be a slight disadvantage in cue saliency for *extrinsic* cues such as the path an object takes. From an ecological perspective such a strategy makes sense as extrinsic properties of an object are less salient with regard to its identity than intrinsic ones (the path an animal takes will be less indicative than the way it moves or its shape). In general, it can be said that this line of experiments speaks in strong favor of a cue integration model of object recognition. More specifically, the cues present in this experiments seem to have been represented independently (virtually no interactions were found between features) as largely orthogonal dimensions in the object representation. Given that participants were not told in advance which properties were going to be manipulated, this also demonstrates the remarkable efficiency with which the visual system is able to extract discriminative information in both the static and temporal domain.

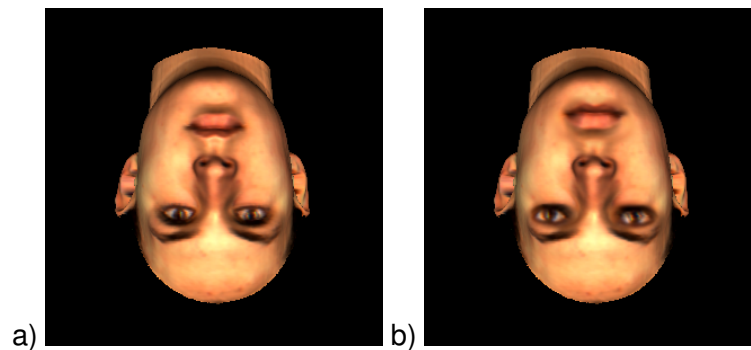
Moreover, the current results demonstrate that the static and motion-based mechanisms mediate object representations that have different characteristics. Specifically, static generalization seems to last over temporal delays but fails across image changes. That is, this static object mechanism supports permanent but spatially refined representations. However, the motion-based mechanism continuously updates information about moving objects. As a result, representations of moving objects might be less spatially refined and rather temporary compared to the representations of static objects.

To summarize, the results from this as well as previous psychophysical and also from physiological studies (see for example, Miyashita [1988]) show that the temporal characteristics of the visual input are an integral part of the object representation and are actively used for learning, recognition and categorization of objects. The question remains, however, how exactly spatial and temporal information are integrated in object representations. In this thesis, I will propose a computational implementation, which provides such an integration as part of the recognition and learning procedure.

1.2.6 Configuration and components

In the previous sections, visual information was - implicitly - viewed in a holistic fashion. Images of objects were treated in a holistic way with each region of the object receiving the same amount

Figure 1.13: Thatcher illusion. The two faces do not seem to be different. When the pictures are viewed upside down, however, the face in b) appears highly grotesque.



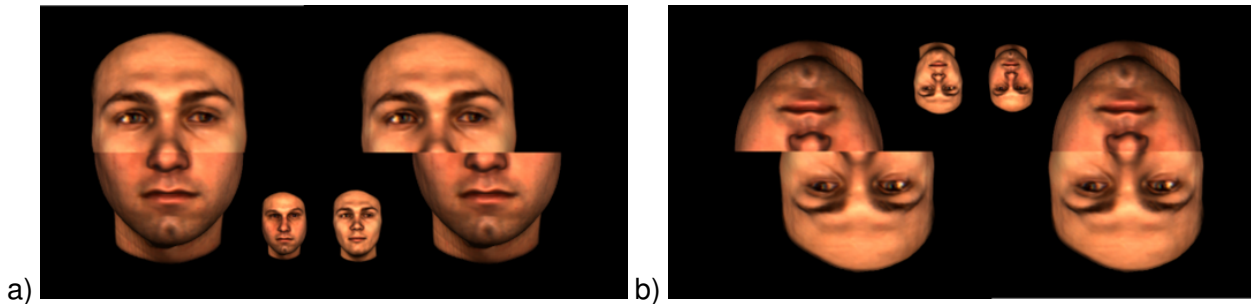
of computational processing. However, as already the fact of attentional processing of the visual world indicates (see especially the results on "change-blindness" [Rensink et al., 1997]), this holistic view might not necessarily be true for object recognition and learning. Indeed, if one considers that objects appear quite often under occlusion or more generally, when parts of the object have been changed, it seems that holistic recognition is not a good recognition strategy. As a solution, processing of objects could be based on *components* - it is important to note that components, are not to be understood as 3D structural parts [Biederman, 1987] but in the sense that in a view-based framework each view can be broken down into *fragments* of local, appearance-based information, which may or may not actually correspond to semantic object parts. Once such components are accessible, an additional information that may be used for recognition consists of the *spatial relations* between these components - this type of object processing strategy can be called *configural*¹⁴. In the following, we present psychophysical evidence from studies on face recognition, which support the view that humans are indeed using a combination of component and configural information during recognition.

It can be said that faces are one of the most relevant stimulus classes in everyday life. Although faces comprise a very homogeneous stimulus class, adult observers are able to detect subtle differences between facial parts and their spatial relationship. According to Bahrck et al. [1975], we are able to recognize familiar faces with an accuracy of 90% or more, even when some of those faces have not been seen for fifty years. In addition, whenever people interact facial expressions are automatically interpreted in order to identify underlying emotional states [O'Toole et al., 2003]. Interestingly, recognition of expressions and identity seem to be two separate capabilities as humans are on the one hand able to recognize any person as smiling - on the other hand they can readily recognize a known person despite sometimes quite intense facial expressions. However, these remarkably adaptive and robust recognition abilities seem to be severely disrupted if faces are turned upside-down [Yin, 1969]. Consider the two pictures in Figure 1.13: Recognizing the depicted identity is more difficult when faces are inverted. Moreover, the two faces seem to have a similar facial expression. Interestingly, if the two pictures are turned right side up, one can easily identify the depicted person and grotesque differences in the facial expression are revealed [Thompson, 1980]. As pointed out by Rock [1973] rotated faces seem to overtax an orientation normalization mechanism such that it is not possible to perceive mentally rotated faces as wholes. Instead, rotated faces seem to be processed by matching parts, which could be the reason why in Figure 1.13 the faces look normal when turned upside-down.

Young et al. [1987] discovered another interesting effect. They created composite faces by

¹⁴Note that this does not mean that no holistic processing takes place. Holistic processing of visual information could still be a separate recognition strategy, which could for example be used for fast and coarse processing of visual information in a snapshot-like manner.

Figure 1.14: Aligned and misaligned halves of different identities. A new identity seems to emerge from the aligned composites (left), which makes it more difficult to extract the original identities. This does not occur for the misaligned composite face (right). It also does not occur for both images if you turn the images by 180°. The insets show the constituent faces.

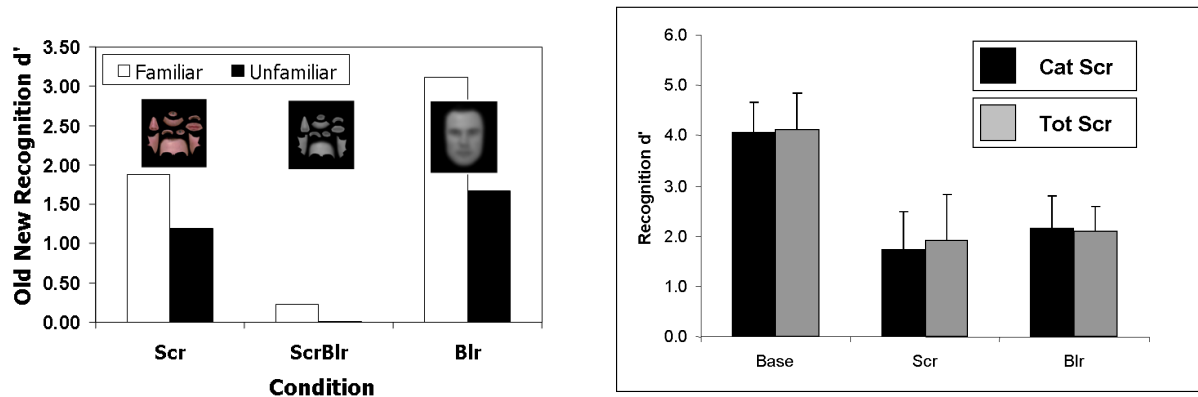


combining the top and bottom half of different faces (Figure 1.14a). If the two halves were aligned and presented upright, a new face resembling each of the two originals seemed to emerge. This made it very difficult to identify the persons from either half. If the top and bottom halves were horizontally misaligned, then the two halves did not spontaneously fuse to create a new face, and the constituent halves remained identifiable. Interestingly, when these stimuli were inverted (Figure 1.14b), the constituent halves of the aligned and misaligned displays were equally identifiable. Furthermore, the participants were significantly faster at naming the constituent halves in inverted than in upright composites. The authors interpret this result as "a dramatic illustration of the absence of interference from configural information in the inverted composites" (p. 753, Young et al. [1987]). When upright, the alignment of face composites creates a new configuration which resembles a new face. When inverted, the processing of configural information seems to be impaired and the two identities are easier to extract based on the facial parts alone.

1.2.6.1 Previous work on configuration and components in faces

The distinction between parts or component information on the one hand and configural information on the other has been used by many studies on human face recognition (for an overview see Schwaninger et al. [2003]). The term component information (or component, piecemeal, part-based information) has been referred to facial elements, which are perceived as distinct parts of the whole such as the eyes, mouth, nose or chin [Carey and Diamond, 1977, Sergent, 1984]. In contrast, the term configural information refers to the spatial relationship between components and has been used for distances between parts (for example, inter-eye distance or eye mouth distance) as well as their relative orientation. There are several lines of evidence in favor of the assumption of component vs. configural representations in face processing. One of the first demonstrations for qualitative differences has been provided by Sergent [1984]. She used pairs of faces that were mismatched either in the eyes or face contour (change of component information) or in the internal spacing of features (change of configural information). The analysis of her results revealed that for upright faces configural and component information were used. In contrast, no evidence was found for the use of configural information in upside-down faces. Note that Sergent [1984] used highly schematic faces that could make it difficult to generalize from this result to the processing of real faces. However, Searcy and Bartlett [1996] found comparable results for colored photographs using different experimental methods. Again, their results suggested that in upright faces component and configural information are used, whereas in inverted faces the processing of configural information is hampered.

Figure 1.15: a) Results from Schwaninger et al. [2002a] showing the dependence of component (scrambled faces, Scr) and configural processing (blurred faces, Blr) for familiar (white bars) and unfamiliar faces (black bars). b) Results from a third experiment in which categorical relations were held intact as opposed to fully scrambled.



a)

b)

1.2.6.2 Experiment 1 - component and configural changes during rotations

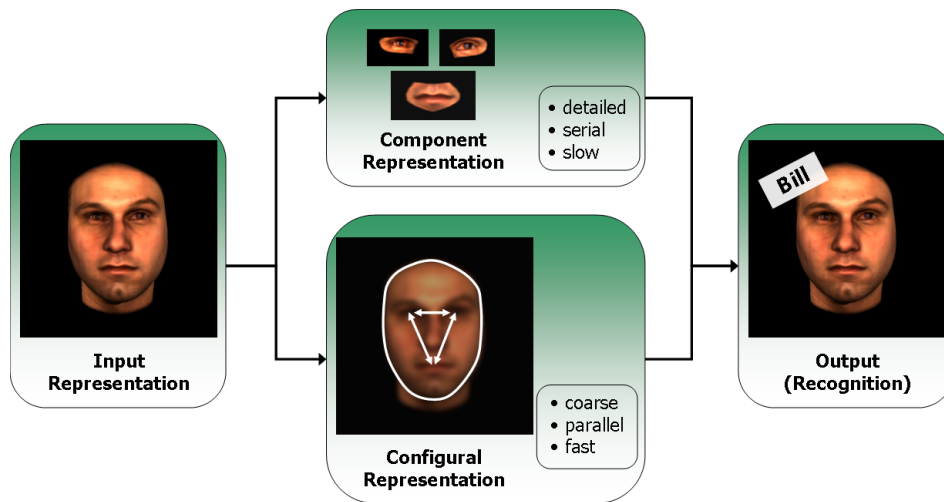
Another demonstration of the differential effects of orientation upon processing component and configural information was found in Schwaninger, Mast, and Hecht [2000]. In their experiment two faces were presented sequentially and the participants had to detect whether components were changed (eyes and mouth replaced) or whether configural information was altered (increased inter-eye distance and distance between the eyes and mouth). Whereas both types of alterations were easy to detect in upright faces, it was difficult for participants to detect configural changes when the faces were rotated. In contrast, the detection of component changes was almost unaffected by rotation. In short, these and other studies provide converging evidence in favor of a qualitative distinction between component and configural information in face processing. However, one possible caveat of studies that investigated the processing of component and configural information by replacing or altering facial parts is the fact that such manipulations are difficult to conduct selectively. Replacing the nose (component change) sometimes alters the distance between the contours of the nose and the mouth and might change the configural information. Similar difficulties apply to configural manipulations when they are conducted by changing the relative position of components. For example moving the eyes apart (configural change) can lead to an increase of the bridge of the nose, which is a component change.

1.2.6.3 Experiment 2 - scrambled and blurred stimuli

Problems like these were avoided in a set of recent psychophysical studies by Schwaninger, Collishaw, and Lobmaier [2002a], Schwaninger, Lobmaier, and Collishaw [2002b]. In contrast to previous studies, they employed a method that did not alter configural or component information, but eliminated either the one or the other. The results of three experiments are depicted in Figure 1.15, where the recognition performance is measured in d' -scores.

In Experiment 1 it was found that previously learned faces could be recognized by human participants even when the faces were scrambled into constituent parts (or components). In these scrambled stimuli configural information was effectively eliminated (Figure 1.15a, left), which makes this result consistent with the assumption of explicit representations of component information in visual memory. In a second condition, parameters of a low pass filter that made the scrambled part versions impossible to recognize were determined (Figure 1.15a, middle). The filter with

Figure 1.16: Integrative model for unfamiliar and familiar face recognition showing the properties of configural and component processing.



this particular set of parameters was then applied to whole faces in order to create stimuli in which local component information would then be eliminated. With these highly blurred stimuli it could then be tested whether configural information was explicitly encoded and stored. It was shown that configural versions of previously learned faces could be recognized reliably (Figure 1.15a, right), suggesting additional and separate explicit representations of configural information.

In Experiment 2 these results were replicated for participants who knew the target faces (white bars in Figure 1.15a). As can be seen from the d' -scores, recognition performance was higher than in the unfamiliar condition. A closer statistical analysis of the data, however, revealed that this increase in performance did not change the relative difference between configural and component information. In other words, this means that the characteristics of processing faces by components and their relations do not change with learning pointing towards a fundamental processing strategy of face recognition.

Finally, Experiment 3 dealt with the type of scrambling. In particular, this experiment addressed the question whether the results would change if the scrambling would preserve the *categorical* relations of the parts, that is, if the facial parts would simply be pulled apart rather than completely scrambled. As Figure 1.15b illustrates, the difference between the two types of scrambling (labelled Cat Scr and Tot Scr, respectively) did not reach significance, which shows that, indeed, scrambling into facial parts results in a consistent processing strategy.

These psychophysical experiments provide converging evidence in favor of the view that recognition of familiar *and* unfamiliar faces relies on component and configural information.

1.2.6.4 Discussion

Based on the results from different psychophysical studies and the model proposed in Schwaninger et al. [2002b, 2003], I want to summarize this line of research in the integrative model depicted in Figure 1.16, which illustrates a cognitive architecture of face recognition. Taking into account that faces seem to be processed in a view-based manner, the input representation should be based on pictorial (or appearance-based) information rather than 3D or structural information (see previous experiments on view-based recognition in section 1.2.2). Based on this visual representation, processing entails extracting local component information and global configural relations in order to

activate component and configural representations (in the brain this could be achieved in so-called face selective areas, Figure 1.16, middle).

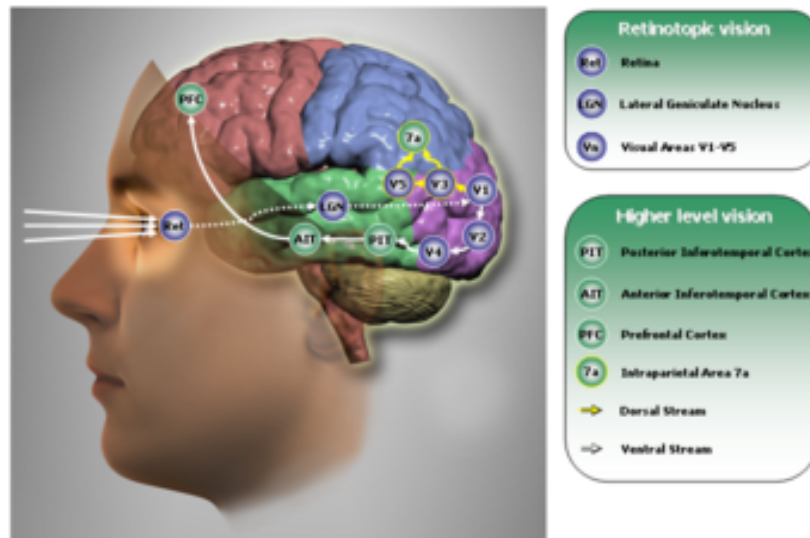
As analyses of recognition times have shown [Schwaninger et al., 2002a], processing of component information seems to take longer than processing of configural information. One of the reasons for this might be that components are accessible only in a serial manner, that is, recognition is done by sequentially matching all components, which requires detailed information to be processed. Configural information, on the other hand, seems to rely on coarse information (see also Goffaux et al. [2005]), which could therefore be processed in parallel and result in the faster matching times observed in Schwaninger et al. [2002a]. In this recent study, repetition priming was used in order to investigate whether the outputs of component and configural representations converge to the same face identification units. Since priming was found from scrambled to blurred faces and vice versa it was proposed that the outputs of component and configural representations converge to the same face identification units (Figure 1.16, right).

In this context I want to stress, that whereas *structural models* predicting view-invariant performance bear little psychophysical plausibility, not all aspects of these models can be dismissed. Instead, this study has actually provided evidence that component information is important for recognition, which is in accordance with several other studies (see earlier discussion on structural models in section 1.2.1). Several important differences, however, exist between structural description models and the model depicted in Figure 1.16):

First, in contrast to the traditional approaches by Marr and Biederman, this model does *not* rely on edge-based representations. Second, the component processes proposed in Figure 1.16 are completely different - both conceptually and computationally - from the geometrical primitives (geons) used in the approaches of Biederman [1987], Hummel and Biederman [1992b]. In contrast to view-invariant geons, component representations are extracted based on pictorial information, which of course is inherently viewpoint dependent. Indeed, there are several lines of evidence against the concept of viewpoint-invariant recognition based on *geons*. For example Tarr et al. [1997] have shown reliable effects of viewpoint for the recognition of objects that were made of 1, 3, or 5 different geon-like parts. Moreover, Hayward and Tarr [2000] have found viewpoint dependent performance for geon-like objects although view-invariant performance would have been predicted for such objects even according to an extended version of RBC proposed by Biederman and Gerhardstein [1993]. Most importantly, Tarr et al. [1998] have revealed that already the processing of one geon is dependent on viewpoint, which accounted for a variety of recognition tasks from sequential matching and matching to sample to naming tasks. Finally, another aspect in which the proposed model differs from structural description models is the number of parts and how they are acquired. According to RBC theory, a small and fixed set of geons possesses sufficient explanatory power to model all relevant aspects of human object recognition. However, human recognition performance relies on many more features, which often are determined by perceptual learning, and are therefore not fixed, but rather dependent on the history of the subject and the task [Schyns et al., 1998]. This provides additional support for the concept of a view-based processing based on a larger number of visual primitives based on pictorial cues.

In summary, it seems that both configural and component information is used for recognition of faces. Given psychophysical evidence on the similarity between face recognition and object recognition in general (see for example, the parallels between view-based face and object recognition results shown in this chapter), it can be expected that these results also generalize to recognition of *other* objects which exhibit an internal structure (for example, animals or artificial objects such as houses and cars) - testing this hypothesis would be an interesting line of future research on structural properties of object recognition.

Figure 1.17: Schematic diagram depicting the flow of visual information in the brain from the retina up to the higher visual areas. Shown are the early vision pathway, the ventral and the dorsal pathway.



1.3 Physiology

This chapter provides a short overview of visual information processing in the brain. The focus here lies more on the broader functional organization of the brain than on detailed modeling of the underlying neural signal processing. Recent methodological and technical advances - especially through functional imaging techniques (fMRI) - have made a more thorough and complex investigation of the brain's processing capabilities possible and in many cases have led to a re-thinking of previously well-established concepts. This chapter tries to account for this development by first presenting the "traditional" and established view of visual processing in the brain and then leading on to some of the more challenging recent findings.

1.3.1 Visual processing in the brain

Visual information in the form of light enters the eyes and is focused onto the retina which contains receptor cells - rods and cones - transforming light energy into neural signals. Cones are responsible for human daylight vision and also form the basis of color vision as the three different types of cones possess different spectral characteristics. In contrast to this, rods are active during low-light conditions and are not sensitive to color. Photo-receptors cluster tightly in a region called the fovea and gradually become sparser towards the visual periphery. This distribution is the basis for a *space-variant processing* of the visual information in brain, where the fovea - which is usually centered on a point of interest in the scene guided by eye-movements or saccades due to shifts in attention - is processed using most of the computational resources whereas the periphery is processed only in a coarse manner. Rods and cones are connected to bipolar cells, which are in turn connected to ganglion cells. These cells represent the receptive fields of the retina, where the size of the receptive fields again varies depending on their retinal position with small receptive fields in the fovea and large receptive fields in the periphery.

Ganglion cells can be classified according to their spatial and temporal response characteristics. Their spatial response is determined by the so-called center-surround organization of the receptive field which is due to excitatory and inhibitory subfields organized into circularly sym-

metric regions. The net effect of this center-surround organization is to enhance contrast sensitivity by reducing the DC component of the visual signal and can be approximately modeled by a difference-of-Gaussians ("Mexican hat") function. In addition, ganglion cells are classified by their temporal response, where the so-called P-cells have a sustained response to contrast in the visual stimulus whereas M-cells have a transient response to contrast. The combination of these two different response characteristics forms the basis of motion perception in the brain. Axons from the ganglion cells constitute the optic nerve which transmits the output of the retina to the lateral geniculate nucleus (LGN). LGN is a layered structure containing retinotopic (that is, a faithful topographical mapping of the retina) maps of the visual field separately for each eye. It projects onto the primary visual cortex (V1), which is the first early visual area of the brain. Interestingly, there are as many forward connections from LGN to V1 as there are backward connections from V1 to LGN. Although the exact role of these connections is yet unclear, it may indicate a massive amount of feedback from higher areas which can be used to influence and bias low-level visual processing (see also next section).

Cells in V1 vary in their receptive field structure and can be classified as simple, complex or hypercomplex cells. Simple cells have a well-defined spatial response characteristics with excitatory and inhibitory regions, which can be well described by a Gabor-type function [Hubel and Wiesel, 1962]. In addition, they exhibit a pronounced orientation selectivity (that is, their firing rate is dependent on the orientation of the visual stimulus) and there exist both monocular and binocular cells. Hubel and Wiesel [1962] have suggested a processing hierarchy where complex cells combine inputs from simpler cell types and hypercomplex cells receive input from both simple and complex cells (see for example, Riesenhuber and Poggio [1999] for a computational model).

V1 is one of the most well-understood brain areas as there seem to be relatively few complications in determining the functions of its cell types (see, however, Ben-Shahar and Zucker [2004]). It projects to many different regions of the brain, whose functions become increasingly complex to model. This difficulty of course ties back to the definition of what exactly constitutes a visual area, where it is sometimes hard to identify common functional properties of a brain area (see Zeki [2003] for a current debate about the division of V3 into two areas). Examples of stimulus dimensions to which cells in V2 and V3 respond include color, shape and motion and thus defy easy classification. It seems, however, that the later area V4 mainly processes color information, whereas V5 (or MT) possesses cells sensitive to different directions of motion and thus is involved in motion processing. Ungerleider and Mishkin [1982] proposed a organizational principle based on a very general processing distinction, which postulates that the dorsal pathway consisting of connections of V1 to the parietal lobe (V1 - V2 - V3 - V5 - MST - 7a) analyzes *where* things are and that the ventral pathway (V1 - V2 - V4 - PIT - AIT) analyzes *what* things are (see also Figure 1.17). This super-structure of visual processing was introduced in order to provide a first breakdown of the types of information being processed by different parts of the brain and became very influential. In the following I will focus mainly on the ventral stream, which in this approach is primarily involved in object recognition and learning.

The final area in the ventral stream is area IT (infero-temporal cortex)¹⁵ in the temporal lobe, which represents the last purely visual area involved in the processing of visual information and is thought to provide the neural basis of object recognition. IT has a number of connections to areas in the limbic system and the pre-frontal cortex as well as a large number of feedback connections to earlier visual areas. Neurons in IT show large receptive fields covering a large portion of the visual field ranging from 1.5° in the posterior area (PIT) up to 30° in the anterior area (AIT). This property was thought to provide an increasing invariance to low-level stimulus parameters (such as size, position, plane rotation), which in turn would be a first step towards global and object specific perceptual encoding rather than local encoding of contrast, motion or color variations. The general idea behind the visual processing hierarchy in the traditional view is thus that later

¹⁵IT is anatomically defined by functional recordings in the *monkey* brain. The human homologue could be seen as area LOC (see below) - this is, however, still under investigation.

visual areas become more and more *invariant* to non-object-specific stimulus properties.

One of the key studies about the functional role of IT regions has investigated the responses of neurons to real-world objects in anesthetized monkeys [Wang et al., 1998]. Although some neurons were found which responded maximally to simple bar-like stimuli, the majority of neurons in PIT preferred complex objects such as star-shapes or circles with protruding elements. Interestingly, neurons were highly sensitive to minuscule changes to these objects such as the relative orientation or thickness of the elements. On the other hand, neurons were quite insensitive to stimulus variations such as size, contrast or retinal location. These findings were taken as evidence that one of the strategies for representing objects might be to use a number of moderately complex visual elements, whose pattern of co-activation encodes the visual appearance of the stimulus. In addition, Wang et al. [1998] found neurons in AIT, which responded maximally to images of whole objects such as faces or cars indicating that already in IT, object specific encodings might be present. In another set of experiments, Logothetis et al. [1994] found AIT neurons, which showed a strong view-based behavior for the paperclip stimuli from Bülthoff and Edelman [1992], whereas again they were invariant to size and location of the stimulus. Their findings provide strong evidence that a neural implementation of view-based object encoding is possible and indeed seems to be used for recognition. As the investigated cells were maximally selective for the holistic stimulus rather than its constituent parts and in addition showed view-selectivity while retaining low-level invariance to size variations, they might be encoding the co-activation pattern of earlier PIT cells and thus form view-tuned units of recognition¹⁶. It is important to stress in this context that an abstraction such as "grandmother" neurons, which specifically encode only one stimulus, does not seem plausible. Rather, the majority of neural responses in this and other experiments showed selectivity for a number of stimuli. A plausible explanation for this finding is that objects are encoded not by a single neuron but by a population code encompassing a number of neurons, which greatly increases the robustness of the representation (see also Wallis and Bülthoff [1999]).

More recent studies have begun to elucidate the nature of object representations in higher brain areas, where especially the role of the ventral occipital-temporal brain area (VOT) has been highlighted. With the help of fMRI (functional magnetic resonance imaging) techniques several regions high up in the ventral stream could be identified, which seem to represent the perceived shape of an object, rather than being sensitive only to low-level form cues. Kourtzi and Kanwisher [2001] used an adaptation paradigm for comparing two stimuli, which differed in a number of manipulations (such as shaded objects versus line-drawings or intact versus scrambled images). As the observed brain activation for similar stimuli gradually decreases due to neural adaptation, an *increase* in neural adaptation in the same brain region can be seen as an indicator for activation of a different neural representation. Indeed, they found that a region known as lateral occipital complex (LOC) showed adaptation for intact and shaded stimuli. Further experiments on 3D depth rotations demonstrated that the VOT seems also to include a view-based representation as adaptation could only be found within a limited range of rotation¹⁷. Interestingly, these results can be related to the single-cell recordings from monkeys done by Logothetis et al. [1994]. Whereas it is still an open question, to what degree single-cell recordings can be related to fMRI scans, which usually look at a million and more neurons, the current results suggest that IT might be using and learning *view-based* object representations.

¹⁶See below for a discussion on how that might relate to the previous section on configural and component processing.

¹⁷James et al. [2002] found reduced view-dependence for familiar 3D objects in a smaller area of LOC called vTO. This result, however, was only obtained for easily identifiable objects, for which view dependence is known to be less (for example, Blanz et al. [1999], Palmer et al. [1981]).

1.3.2 Beyond the traditional view

Here I want to briefly present studies that challenge some of the conclusions drawn from the experiments described above and which may lead to a better understanding of the functional organization of visual processing in the brain.

First of all, returning to area IT and the study by Wang et al. [1998], it can be noted that their particular choice of stimuli in the experiments seems to be quite arbitrary. Whereas it might be still true that IT represents objects by a number of visual primitives, several studies have failed to show conclusive evidence for a set of stable and general primitives. Rather, it seems that if such a set can be found it is highly task-dependent, which points at a more general processing function of IT. Indeed, Op de Beeck et al. [2001] reported recognition experiments with visual stimuli whose complex outline shape was generated by three sets of parameters and thus could be easily described in a low-dimensional space. Correlating psychophysical results on how humans grouped these stimuli into sets with recordings of IT neurons' responses for monkeys doing the same task they found good agreement for a low-dimensional representation in both conditions. This set of experiments shows that IT neurons seem to be capable of extracting and representing important stimulus properties in a low-dimensional pattern space. Rather than looking for a global and fixed set of visual primitives this study is one of the first to investigate the *processes* behind the formation of object representations and the *learning capabilities* of this area (see also Sigala et al. [2002] for a study on categorization of familiar shapes).

In a recent study, Baker et al. [2002] investigated these learning capabilities more closely by looking at how training influenced neural responses to holistic and part stimuli in monkey IT. Stimuli consisted of simple shapes, which were derived from four different end-parts joined by a line. After training, single-cell recordings were used in order to determine neural responses to learned and novel stimuli. Their main result was that learned stimuli showed greater selectivity for both configuration and holistic processing, where the holistic contribution was larger than the sum of the two part contributions. These results show interesting parallels to developmental studies on the role of holistic and configural processing [Schwarzer and Massaro, 2001] and also confirm the notion of two explicit routes of object processing as discussed in the previous section (see also Figure 1.16).

Another recent fMRI study investigated the role of earlier visual areas in shape perception. Traditionally, the integration of local shape elements into global shapes was thought to reside in higher occipitotemporal areas. Using contours defined by Gabor patches as stimuli, however, Altmann et al. [2003] found that *both* early and higher visual areas exhibited stronger responses to contours defined by local elements grouped by orientation and/or disparity than to randomly oriented local elements. Interestingly, both the behavioral detection performance of participants *and* the fMRI signals decreased for disrupted contours, whereas they increased again when additional cues such as disparity were used to group misaligned contour elements. This suggests that similar neural mechanisms could support grouping of local elements to global shapes by such different visual features as orientation or disparity. These findings are consistent with the view that recurrent mechanisms of visual processing mediate global shape perception by feedback connections.

The traditional view keeps the two processing pathways largely separate with the ventral pathway processing the "what" and the dorsal pathway processing the "where" information [Ungerleider and Mishkin, 1982]. However, even at first glance it is clear that these types of information are not completely separable, as interaction with the environment requires the convergence of location and identity for action. For this reason, a strict separation of the two pathways based on such a broad functional description as in Ungerleider and Mishkin [1982] does not seem to be plausible. As neurophysiological studies show, there is indeed a large number of connections between the parietal and temporal lobes, which could serve as a means of communication (see for example, Bullier [2001], Zhong and Rockland [2003]). In a recent fMRI study by Kourtzi et al. [2002] the influence of motion processing areas on object recognition was investigated, thus directly looking at the influence of the dorsal on the ventral pathway. Stimuli consisted of intact or scrambled images

of everyday objects and were presented stereoscopically. The object-selective fMRI responses that were observed for shaded objects suggest that binocular disparity-tuned neural populations in MT/MST are not only involved in the analysis of local disparity signals but may also be engaged in the processing of the perceived 3D shape of objects independent of the cues (binocular or monocular) defining the object structure. Consistent with this interpretation, the strongest object-selective responses in MT/MST were observed for objects defined by shading cues that specify 3D structure than for objects defined by motion or stereo that simply facilitate segmentation of objects from their background. Thus it seems that in contrast to traditional theories of visual processing, their findings suggest strong functional interactions between the neural mechanisms involved in the processing of shape and motion information about objects - localized in this particular case in the area MT/MST which is directly involved in the processing of shape properties of both moving and static 3D objects. This study could lead to further insights on how *spatio-temporal* representations of objects might be encoded and recognized in the brain (see previous section).

1.4 Conclusion

Recent developments in cognitive sciences have dramatically reshaped our understanding of visual processing both from a psychophysical and neuro-scientific perspective. Object recognition is not a matter of faithful reconstruction of 3D structures but seems to rely on transformation-dependent processing of spatio-temporal, view-based representations. In a more extreme view one could say that a complete invariance to any kind of stimulus property or viewing condition does neither seem to exist on a psychophysical nor on a physiological level. However, this does not mean that view-based theories in their current form can explain all of the experimental data. Indeed, RBC theory has undergone several iterations in the form of a computational abstract model developed by Hummel and Biederman [1992a], Hummel [2002] - for example to include view-based processing routes for fast recognition whereas view-invariant (geon) descriptions are available as a time-consuming processing route¹⁸. Additional evidence for the extraction of non-accidental (and thus largely invariant) properties, which form the basis of the geon definition, has come from psychophysical and physiological studies, where changes in non-accidental properties were more easily detectable than changes in metric properties (such as size) of an object [Vogels et al., 2001]. On the other hand, Edelman and Intrator [2001] have taken up the notion of structured representations and developed a computational framework, which is based on an implicit structural description in a view-based interpretation. Their biggest criticism of RBC theory concerns the apparent problems of both Hummel and Biederman [1992b] and indeed the computer vision community as a whole in presenting a computational implementation of RBC theory, that is capable of working with and interpreting of *real-world input images*.

Interestingly, a recent psychophysical experiment by Foster and Gilson [2002] has taken a more integrative standpoint by investigating to what extent a *combination* of view-based and view-invariant processing might explain object recognition. Their experiments employed the same inter-extra-ortho paradigm using novel paper-clip like stimuli in a same-different task as in Bülthoff and Edelman [1992]. The stimulus parameters that were varied, however, now were metric properties (thickness or size of the paperclips) or non-accidental structural properties (the number of limbs/parts of a paper-clip). Results for *both parameters* showed strong viewpoint dependency for an angular range of 90° around the learned viewpoint whereas for the remaining angular range, the viewpoint dependency was much less pronounced. The only difference between the two conditions were that non-accidental changes were faster and more accurately recognized (in accordance with the above mentioned findings) with an otherwise completely identical angular dependence pattern. The authors interpreted their findings as the result of two processes which

¹⁸One of the advantages of a geon description might lie exactly in their "unspecific" responses to exemplars, which would enable categorization processes rather than recognition of exemplars. However, it is exactly categorization, which was shown to work extremely fast and thus conflicts with the assumption of time-consuming processing for geons.

are additively combined in order to yield the final recognition result across depth rotations: the first process is independent of viewpoint but dependent on structure, whereas the second process is dependent on viewpoint but dependent on structure. Note that this means a departure from the traditional view of view-based and structural processing towards a new combination of view-invariant processes.

Thus, instead of taking the extreme standpoints of view-based versus view-invariant processing one might envisage a visual processing framework in which features are selected according to the current task, where the optimality, efficiency and thus the dependency on viewing parameters of the features depend on the amount of visual experience with this particular task. In this context, it is important to stress that a good definition of what constitutes a (visual) feature both from an abstract as well as from a functional (neural) level still has to be developed. Here, a close collaboration between a number of disciplines including neuro-physiology and psychophysics as well as machine learning could offer new ideas based both on experimental and theoretical work.

Further proof for the complexity of object recognition processes comes from recent functional imaging studies, which show that while performing object recognition tasks a large number of areas is recruited ranging from motor-related areas to motion-processing areas and from high-level visual areas to early visual areas. This is strong evidence for a highly connected recurrent network in which object recognition is achieved only through the tightly integrated communication between areas providing visual, motor action and other kinds of behaviorally relevant processing expertise (further examples might be areas responsible for auditive, tactile or proprioceptive processing). The realization that object recognition is perhaps not so much relying on extraction of *invariant* visual features but rather on *appearance-based, spatio-temporal and inherently multi-modal representations* might be a first step towards a further understanding of the processes behind our astonishing recognition performance. The focus of research could thus be shifted from the investigation of isolated phenomena towards the properties and functional organization of the (dynamic) interaction of the different modalities.

Even though the results presented in this chapter seem to produce more questions than answers, they nevertheless contain some tangible and concrete aspects, which can be used as the basis for a biologically plausible framework of recognition and in addition can provide the starting point for a computational implementation of object recognition. In summary, these are:

- object learning and recognition involves spatio-temporal and multi-modal representations
- object recognition relies on transformation-dependent processes
- object representations can be based on pictorial, image-based features from a number of cues
- one of the assumptions that enables such pictorial representations to be learned is temporal continuity
- object recognition is structured by a combination of holistic and configural processing in a view-based context

In chapter 3, I will use these basic assumptions to define a computational recognition framework where the main goal is not to focus on a possible neural implementation of the system but rather on functional organization aspects of object recognition. The final justification for this approach will be given both by modeling of psychophysical data and by computational recognition experiments. Taken together these experiments will demonstrate that integration of results from perceptual research can be used to build a system with increased levels of performance compared to more monolithic approaches.

Chapter 2

Computational approaches to object recognition

Whereas the previous chapter focused on the cognitive basis of recognition, in this chapter I will focus on computational object recognition within the field of computer vision - both of these chapters taken together will form the basis of the computational framework introduced later.

In relation to object recognition, the overall goal of computer vision in general could be defined as "building a computational system that can understand a scene". This definition of course entails a much broader - even cognitive - perspective of processing of visual input and, for example, includes analysis of categories, higher-level object relations, scene context, etc. Nevertheless, object recognition is a fundamental prerequisite to achieve this ultimate goal of scene understanding that can be traced back to the earliest days of the field of computer vision [Marr, 1982].

The second context with which I want to connect this analysis is the field of machine learning. This research area is rooted in mathematics and in its broadest sense is concerned with the analysis and processing of data in order to solve encoding, recognition or regression tasks.

In the context of computer vision and machine learning one could formulate a simple, generic framework of object recognition in three steps: first, input data is used in a supervised manner to build object model(s). Here, the term supervised learning is defined in its broadest sense, which encompasses for example choosing input data (which images does the machine learn?), supplying pre-labeled data (which object is present in each image?), but also including a-priori models of data formation (for example, images only show faces undergoing a depth rotation on the horizontal axis). Second - and this will be the focus of this chapter - learning of object models is done in a two-step manner by extraction of features from the input data and further processing of these features by a learning algorithm. Such an algorithm might for example exploit statistical redundancies in the data to yield a sparser representation of the given objects. The first two steps are usually done off-line with a given database and thus result in a number of pre-determined object model(s). The third step is the online or recognition phase, where again features are extracted from the input data to yield a compatible data representation. The learning algorithm then compares this representation with the stored object model by means of a suitable distance (or similarity metric) and decides whether the input data could be recognized or not. The recognition phase is thus mostly unsupervised as extraction of the data representation and classification proceed without explicit intervention by an outside agent.

The two important elements in this framework from an implementational point of view are thus

- the choice of the feature extraction method for representing data and
- the choice of the learning and classification algorithm for encoding and recognizing these representations.

In the following, I want to provide a brief overview of previous work on computational recognition frameworks following two main lines. The first essentially provides a coarse chronological ordering

by focusing on data representation, whereas the second focuses on classification algorithms used during learning and recognition.

2.1 Data representation

The following discussion on data representation is presented in the context of *cognitive* object recognition models as presented in chapter 1. The goal of feature extraction is thus defined as providing a data representation that is suitable for efficient recognition of visual data for the human visual system. In computational terms, the input to an algorithm consists of a number of images, which are usually given as an array of pixels. The two categories of data representations I want to discuss in this context are *structured shape models*, which extract 3D shape from images and *appearance models*, which extract statistical information about the 2D intensity distribution of pixel values. These two categories have their natural counterparts in the structural and view-based model mentioned in the previous chapter.

2.1.1 Structured shape models

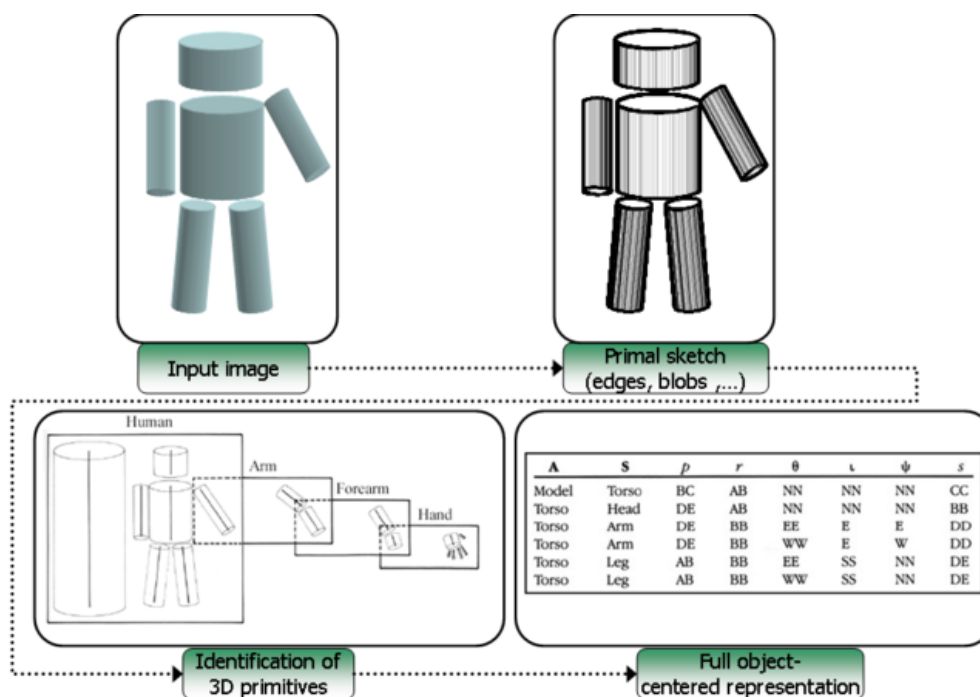
The main goal of structured shape models is to recognize objects by their 3D shape, which entails a more or less detailed 3D object representation that has to be extracted from 2D images. It should be noted that in general the reconstruction of 3D shape from *a single* 2D image is an ill-posed problem as an infinite number of perspective projections give rise to the exact same retinal image. A solution can be found by introducing constraints in order to make this problem tractable, which - depending on the type of constraints - leads to a number of frameworks known as “shape from X” in the literature. One of the most obvious sets of constraints, which is available for the human visual system, is binocular analysis (“shape from stereo”) of two images taken from the slightly different viewpoints of the two eyes. The two different viewpoints result in different retinal positions for a given point in 3D space (“disparity”) - finding corresponding points in the two images then enables a 3D reconstruction of the scene. The same line of argument holds for image *sequences*, where “shape from motion” can be used to infer 3D structure from several subsequent images (“motion parallax”, “structure from motion”). Other monocular methods used by the visual system such as “shape from shading” generally require heavy constraints in the form of prior knowledge in order to provide unique solutions. It is, however, clear that the brain has the machinery to derive 3D information even from single images, which provides the starting point for models discussed below.

Starting with Marr and Nishihara [1978] many systems were developed to reconstruct 3D structure from input images (for example, Lowe [1985], Beardsley et al. [1996]; for an early overview of techniques see also Faugeras [1993]). Although they were able to perform very well under specific conditions - such as computer generated simulations of simple geometric worlds or indoor environments with restricted sets of objects - the general reconstruction quality and with that the level of object recognition performance of these systems fell short of unconstrained object recognition in humans.

2.1.1.1 High-level shape models

David Marr was the founder of today’s field of computer vision by stating that vision should be studied from a computational perspective - from an algorithmic point of view. Starting with the approach proposed by Marr and Nishihara [1978], he and his colleagues developed a hierarchical, high-level framework of the visual system (see Figure 2.1). This research was motivated by what was known at the time of the functional organization of the human visual system. The first steps in this framework involved extraction of the so-called primal sketch from an input image - this sketch included several low-level visual primitives such as edges, blobs, etc. The next stage consisted

Figure 2.1: Structural description of objects according to Marr and Nishihara [1978].



of fitting 3D primitives to these visual primitives in order to extract a 3D description of the visual object. The final representation of objects was based on hierarchical arrangements of suitable 3D primitives which can be used to approximate the shape of objects - so-called generalized cylinders. A cylinder corresponding to the main axis of an object forms the first level of the hierarchy. The locations and orientations of cylinders in the next level are specified relative to this cylinder. Each of the cylinders in this level serves as a reference point for cylinders at the next hierarchical level thus creating a tree-like representation. Since the position of each part is defined relative to other parts of the object, the description of an object's shape will be the same regardless of viewpoint. As already pointed out in the previous chapter, the main problem with this model has proved to be the robust extraction of visual primitives, which are able to support the subsequent analysis by generalized cylinders.

2.1.1.2 Alignment to 3D representations by non-accidental properties

A more specific way of extracting features was proposed by Lowe [1985, 1987] (see also Figure 1.3a). In this model, objects were represented as 3D wire-frame-like shapes. Lowe formulated the problem of recognition in a reverse fashion: given a 3D model of an object and corresponding points in the image and on the object model, what are the transformation parameters such that the resulting projection of the model fits the given points in the image. This was one of the first approaches which addressed the problem of indexing into stored object representations for recognition. Since a full comparison of all image features against all models would be prohibitively expensive, the implementation was based on salient features in the image, which could be used to select stored representations. The human visual system was known to automatically group visual primitives in an image by certain properties such as collinearity, coterminality, symmetry and smooth continuation - principles discovered by Gestalt psychologists in the early 1920s and termed non-accidental properties by Lowe [1985, 1987]. The system thus sought to explicitly extract such properties from edge images of scenes and use these to recognize objects and was one

Figure 2.2: Formation of geons (from the original study Biederman [1987]).

Partial Tentative Geon Set Based on Nonaccidentalness Relations







Geon	CROSS SECTION			
	Edges Straight S Curved C	Symmetry Rot & Ref ++ Ref + Asymm -	Size Constant ++ Expanded - Exp & Cont -	Axis Straight + Curved -
	S	++	++	+
	C	++	++	+
	S	+	-	+
	S	++	+	-
	C	++	-	+
	S	+	+	+

Figure 7. Proposed partial set of volumetric primitives (geons) derived from differences in nonaccidental properties.

of the first successful computer vision systems applied to a larger number of real-world images.

Non-accidental properties were later taken up as the basis for extraction of another set of 3D primitives - the so-called geons by Biederman [1987]. In this influential theory of object recognition, Biederman claimed that objects are represented by a small set of these geons (24, in later versions of the theory also 36 - see Figure 2.2), which are 3D geometrical shapes generated by a hierarchical set of rules and which give rise to non-accidental image features. A full object representation was then constructed by combining different geons - each of which represents a distinct part of the object. The recognition by components framework proved to be able to explain many psychophysical results on categorization of objects and was developed into a high-level recognition network by Hummel and Biederman [1992a], Hummel [2002].

Again, however, both of these models rely on robust extraction of non-accidental properties from visual data, which up to now no computer vision algorithm has been able to deliver with satisfactory performance. It seems that a good interface between the high-level components of the models (be it geons or generalized cylinders) and the low-level visual input (pixels) is still lacking in order for these models to interpret natural scenes.

2.1.1.3 2D view alignment

Ullman and Basri [1991] realized that storing 3 or more views of an object along with full correspondence between all points (so-called alignment, see Figure 1.3b) would be equivalent to storing a 3D model of the object. This view-based approach to reconstruction and recognition was based on a theorem from earlier work by Ullman [1979], which states that 3 or more (orthographic) 2D views of 4 non-coplanar points that maintain fixed 3D inter-point distances (that is, rigid objects), uniquely determine the 3D spatial structure of those points. Thus, it was not necessary to store the full 3D model, but only a small number of suitably selected views - with the only constraint that 4 points should be visible in all of these views. Given a number of corresponding points in the image and in all of the views, a novel view N is simply determined by a linear combination of n stored views M_i , such that

$$N = \sum_i^n \alpha_i M_i$$

Conversely, a novel view N is of the same object M if the above relation holds. Although this approach does not explicitly rely on a 3D reconstruction of the object, it nevertheless fits the category of “structured shape models” quite well because of the equivalence of the multiple views to the underlying 3D model.

The main difficulty with this approach in terms of a computational implementation has proved to be the reliable extraction of corresponding points between *all* images. In addition, the required number of corresponding points has to be quite large in order to robustly *recognize* exemplars within a given category as two exemplars might only differ by a few features which would have to be captured in all stored training views.

2.1.2 Statistical appearance models

The main characteristic of statistical appearance models is that they do not rely on a 3D data representation but instead extract features directly from 2D images and therefore describe the visual appearance of the object. In general one could cluster these approaches into two main categories: *global approaches*, which use the whole image in order to extract a data representation, and *local approaches*, which focus on a subset of image regions.

2.1.2.1 Eigenobjects

Principal Component Analysis (PCA) is one of the most influential techniques for appearance based recognition of objects. It was applied for the first time by Kirby and Sirovich [1990] to recognize images of human faces. This was followed by the development of Eigenfaces - a term coined by Turk and Pentland [1991] - and later developed into a general learning strategy for recognition of 3D objects through a number of views [Murase and Nayar, 1995]. In all of these approaches, PCA is used to project the high-dimensional image data into a low-dimensional feature space, which is spanned by the eigenvectors of the correlation matrix of the images. The general equation for this correlation matrix of n images \vec{x} is:

$$C = \frac{1}{n} \sum_{j=1}^n \vec{x}_j \cdot \vec{x}_j^T$$

The principle components (that is, the eigenvectors \vec{P}) of C can now be found by solving the following eigenvalue problem:

$$\lambda \vec{P} = C \cdot \vec{P}$$

In a statistical sense, PCA finds orthogonal dimensions of maximum variance in the data and can be visualized as a simple coordinate transform into a variance-based coordinate system. The next step, which is usually performed in PCA approaches for object recognition, is to *reduce* the number of retained eigenvectors according to a threshold applied on the eigenvalues. In a typical application, the eigenvalues decrease sharply after a few components such that data reduction without loss of statistical power is possible. In order to deal with the high dimensionality of images, a number of approaches have been suggested. One of the most commonly used approaches employs a simple trick by applying PCA not on the correlation matrix between all *pixels* but on its transpose C^T .

One of the major drawbacks of PCA is that it is only a linear algorithm - it essentially assumes Gaussian-distributed data, which might not always a good approximation of the feature space. Recent approaches addressing this problem extend PCA to nonlinear feature spaces [Schölkopf and Smola, 2002] or - as ICA [Hyvarinen and Oja, 2000] does for example - separate the data according to higher statistical moments.

2.1.2.2 Histograms

In general, histograms for the case of object recognition represent a *discretized* distribution of filter responses or image features within an image in order to form a low-dimensional feature vector. One of the first successful approaches using histograms for recognition of objects was proposed by Swain and Ballard [1991] and was based on histograms calculated across RGB color channels of an image. Given the small size and simplicity of the histograms, they found a surprising robustness to changes in orientation, scale, partial occlusion and changes in viewing angle. One disadvantage from a conceptual point of view is that such histograms are invariant against permutation of pixels in the image, which is certainly not a desirable effect for modeling human perception of images, as it is severely affected by scrambling. Additional drawbacks for color histograms are sensitivity to lighting conditions and that for many object classes color is not a discriminative feature. Nevertheless color histograms have proved to be remarkably robust in a variety of recognition tasks despite their conceptual simplicity.

Schiele and Crowley [2000] proposed a general histogram framework for recognition by introducing multidimensional, receptive field histograms to approximate the probability density function of local appearance. Their recognition algorithm calculates probabilities for the presence of objects based on a small number of vectors of local neighborhood operators, such as Gaussian derivatives at different scales. The method obtained good object hypotheses from a database of 100 objects using a small number of vectors and was extended to work for even larger databases.

2.1.2.3 Local features

The extraction of local features, or interest points, from images has a long tradition in computer vision. Based on early work on stereo by Moravec, suitable detectors for such interest regions have been developed and refined (see, for example, Harris and Stephens [1988], Schmid et al. [2000]).

Based on such local features, Schmid and Mohr [1997] developed a system that could recognize objects in the case of partial visibility, image transformations and even within complex scenes. The approach was based on the combination of differential invariants computed at interest points with a robust voting algorithm and semi-local constraints. Recognition was based on the computation of the similarity (represented by the Mahalanobis distance) between two of such feature vectors. This was one of the first systems that was tested on a larger database and achieved excellent recognition results.

The idea of computing local features was further developed by many authors in order to include invariances (such as viewpoint invariance and affine invariance [Schaffalitzky and Zissermann, 2001, Mikolajczyk and Schmid, 2002, 2005], scalespace selection [Laptev and Lindeberg, 2003]). One of the recent recognition frameworks in this spirit is the work by Lowe [1999, 2000, 2004] who proposed scale invariant feature descriptors based on high-dimensional derivative histograms. The most prominent features of this approach are its inherent scale invariance and the excellent performance characteristics which are a result of highly discriminative feature descriptors and optimized matching strategies based on Hough transforms and pose verification (see also chapter 6).

2.1.2.4 Multiple features

Up to now, most approaches mentioned have only looked at one particular data representation. From the study of the human visual performance, however, it seems clear that the visual system has access to a large variety of different features (color, binocular vision, shading, etc.), which give rise to different data representations.

In this context, I want to mention the work of Mel [1997], who - based on this idea of multiple data representations - collected a number of different low-level features including edge strength,

image response to a number of simple orientation detectors and textural properties. These features were modeled after simple, biologically plausible low-level features of the human visual system. The responses from these features were evaluated over the whole image and concatenated to form a single, large feature vector. Recognition performance with this collection of features was surprisingly high and demonstrated the power of integration of multiple features. In addition, the results showed for the first time that object recognition was even possible for *deformable* objects under limited non-rigid transformations.

One open problem that also interfaces with ongoing perceptual research (see Ernst and Banks [2002] and chapter 8) is how to integrate the different data representations in an optimal way. This question has given rise to the field of "sensor fusion" in which efficient strategies for combining different representations are developed.

2.1.2.5 An example: Computational approaches to face recognition

Face recognition is certainly one of the most active topics in computer vision. Within this field, there has been a steady development of algorithms suitable for detection and recognition of faces in various conditions. With the introduction of the standardized FERET database and test procedure (Phillips et al. [2000]) these algorithms can now be evaluated under controlled and reproducible conditions. Interestingly, computational approaches to face recognition have developed historically from simple, geometric measurements between sparse facial features to appearance-based algorithms working on dense pixel data. Although it was shown that recognition using only geometric information (such as distances between the eyes, the mouth, etc.) was computationally effective and efficient, robust and automatic extraction of facial features has proved to be very difficult under general viewing conditions (see Brunelli and Poggio [1993]). In the early 1990s, Turk and Pentland [1991] developed a different recognition system called "Eigenfaces", which used the full image information to construct an appearance-based, low-dimensional representation of faces. This approach proved to be very influential for computer vision in general and inspired many subsequent recognition algorithms. Although these algorithms were among the first to work under more natural viewing conditions, they still lacked many important generalization capabilities (including changes in viewpoint, facial expression, illumination and robustness to occlusion). A recent development has therefore been to extract local, appearance-based features, which are more robust to changes in viewing conditions. Some examples of these approaches are: graphs of Gabor Jets (Wiskott et al. [1997]), Local Feature Analysis with PCA (Penev and Atick [1996]), image fragments (Ullman et al. [2002]) and interest-point techniques (Schiele and Crowley [2000], Lowe [2004], Weber et al. [2000b]). Going beyond these purely 2D approaches, several approaches have been suggested that use high-level prior knowledge in the form of detailed 3D models in order to provide an extremely well-controlled training set (most notably, Romdhani et al. [2002], Weyrauch et al. [2004]). Recently, there has been growing interest in testing the biological and behavioral plausibility of some of these approaches (for example, O'Toole et al. [2000], Wallraven et al. [2002]). However, the work done in this area so far has focused on comparing human performance with a set of "black-box" computational algorithms.

In this thesis, I want to go one step further by proposing to look at specific processing strategies employed by humans and trying to model human performance with the help of the computational architecture proposed in the following chapter. This will not only allow us to better determine and characterize the types of information humans employ for face processing but also to test the performance of the proposed architecture in other recognition tasks.

2.2 Classification algorithms

This part focuses on the classification algorithms used to process object representations during recognition and learning. As an exhaustive overview of the field is outside the scope of this the-

sis (see for example, Duda et al. [2001], Schölkopf and Smola [2002] for excellent reviews and textbooks on this topic), I want to focus on a few representative algorithms that have been applied to object recognition in computer vision before. The discussed algorithms are supervised, that is, they require as training input a labeled database, and in addition they are sparse classifiers, that is, the final classification decision given a new test exemplar is based only on a subset of the training data. The process of classification can thus be represented by a decision function, which, given a test exemplar, yields the class label of the appropriate class. An equivalent representation of the decision function is given by the so-called separation hyperplane, which in the case of a simple linear decision function between two classes of 2D data can be visualized as a *line* separating the two classes.

The notation for the following sections is as follows: the input database consists of c classes, each of which contains $n(c)$ data vectors \vec{x}' . As the discussed algorithms are supervised, each data vector contains not only the data, but also the class label: $\vec{x}' = \{\vec{x}, y\}$, $\vec{x} \in \mathcal{R}^m, y \in \{1, \dots, c\}$. The decision function for a pattern becomes thus $f : \mathcal{R}^m \rightarrow \{1, \dots, c\}$, $f(\vec{x}') = y$.

2.2.1 K-means with n-nearest neighbor

K-means together with a n-nearest neighbor decision function represents the most basic classification algorithm in computer vision. In this context, k-means is used as the learning step of the classification algorithm, whereas n-nearest-neighbor is used to define the decision function. In this framework, a labeled class is represented by k prototypes and new test data are classified according to the label of the n nearest neighbor.

If $k = n = 1$, we have the simplest setup of this classification scheme, the so-called nearest-neighbor prototype classification, where each class is represented by its center-of-mass. Prototypes are thus calculated as the means of each class:

$$\vec{\mu}_c = \frac{1}{n(c)} \sum_{i=1..n(c)} \vec{x}_i$$

Considering the Euclidean norm as the distance metric, we can then write the decision function simply as:

$$f(\vec{x}) = \operatorname{argmin}_c \|\vec{x} - \vec{\mu}_c\|^2$$

Prototype classification is maximally sparse, as each class is represented by only a single vector. Given a new test vector, the decision function in this case looks for the most similar prototype and assigns its class-label to the test vector. It seems obvious that this classification scheme yields only a very crude representation of the data, which might occupy rather complex manifolds in high-dimensional feature space. This classifier is thus often used as a basic “benchmark” for recognition tasks to establish a baseline classification performance.

If $k = n(c)$ and $n = 1$, that is, when the number of prototypes equals the number of class members, the algorithm degenerates to a “learning-free” nearest neighbor classifier. This represents the other extreme end of a classification scheme with *all* trained elements being stored in the final object model, which is also often used as a baseline classifier in the object recognition literature. The reason for this is mainly that with this classifier one can study the metric and the data representation in more detail. The most obvious drawback of this technique are its prohibitive storage requirements which allow efficient use only for a small number of classes and low-dimensional data representations.

If $k = n(c)$ and $1 < n \leq n(c)$, the decision function becomes

$$f(\vec{x}) = \operatorname{argmin}_c \sum_{j=1}^n \min(j)_c \|\vec{x} - \vec{x}_c\|^2$$

where the function $\min(j)_c$ finds the j th minimum value in each class c . Using an n -nearest neighbor decision function, the final class label depends on the distance of each test vector to the n nearest data vectors of each class. The main problem with this strategy is how to determine the value of n , which often can only be done by post-hoc analyses.

If $1 < k < n(c)$, an iterative clustering algorithm [Duda et al., 2001] is used to compute more than one “prototype” per class, which therefore represents an extension of the single prototype classifier. This k -means algorithm has two main difficulties: first, the iteration is rather sensitive to initialization of the means and thus might not converge to a global optimum. Second, it is often not clear how k should be chosen in order to faithfully capture the high-dimensional feature space.

In summary, the k -means (with n -nearest neighbor classification) algorithm represents a simple and straightforward classification scheme. Apart from the problem of how to select the crucial parameters k, n , another drawback is that the processing time of all of the above nearest neighbor techniques is of the order of $O(k \cdot n)$ - which is especially problematic in the “learning-free” case of $k = n(c)$.

2.2.2 Radial basis function networks

In 1990, Poggio and Edelman [1990] proposed Radial Basis Function (RBF) Networks for recognition purposes which were motivated by function approximation theory. In this framework, each node (neuron) represents the learned input in feature space by a Gaussian function centered on the input value. In the case of object recognition the output of the network will converge to a single value with the following decision function:

$$f(\vec{x}_i) = \sum_{k=1}^N w_k \exp(-\|\vec{x}_i - \vec{\mu}_k\|/\sigma_k^2)$$

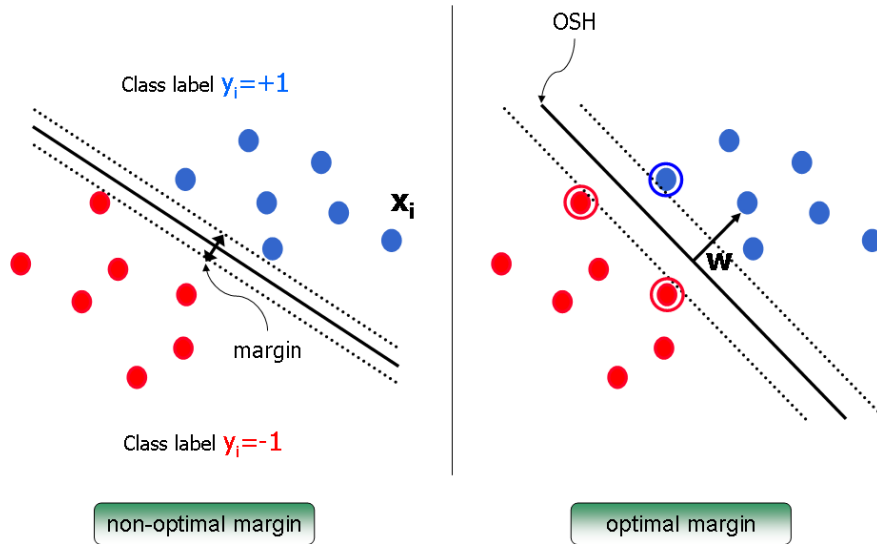
Standard algorithms such as Singular Value Decomposition can be used to estimate the parameters μ, σ of the Gaussians and the weights for the output summation from the training data. The ensemble of Gaussians is thus used to approximate a full continuous object representation by a small set of training data. As with the previous examples, choosing the number of neurons in an RBF network can probably be only solved using heuristics (see for example, Bischof and Leonardis [2001] for a framework that optimizes the number of neurons based on MDL principles and also provides an online learning example). If applied to recognition of 2D image data, the RBF framework is characterized by view-dependent performance for recognition of new input values, where slightly different values can still be recognized by the network due to a large enough support of the Gaussian function [Poggio and Edelman, 1990] (see also Figure 1.3). The psychophysical experiments involving computer-generated unfamiliar objects (“paperclips”) conducted by Bühlhoff and Edelman [1992] could successfully be modeled using RBF networks. RBF networks represent an approach to view-based recognition in which recognition is done by view-interpolation in feature space and thus were also proposed as a biologically plausible framework for modeling human object recognition performance.

2.2.3 Support vector machines

In this section I want to give a brief overview of the principles underlying the Support Vector Machine (SVM) framework, which has received much attention in the literature as it is firmly rooted in statistical learning theory. Here, I want to concentrate on *binary* classification with SVMs - for further details and the extension to multiclass settings see discussion in chapters 6 and 7 as well as Cristianini and Taylor [2000], Schölkopf and Smola [2002], Vapnik [1998].

Consider the problem of separating the set of training data $\{\vec{x}'_i\}_{i=1}^c = \{(\vec{x}_1, y_1), \dots, (\vec{x}_c, y_c)\}$ into two classes, where $\vec{x}_i \in \mathbb{R}^m$ is the feature vector and $y_i \in \{-1, +1\}$ its class label. If we assume that the two classes can be separated by a hyperplane $\vec{w} \cdot \vec{x} + b = 0$, and that we have no prior

Figure 2.3: The figure illustrates the concept of the margin, which is maximized (left: non-optimal margin, right: optimal margin) in order to obtain an optimal separation of the two classes. The optimal margin defines the optimal separation hyperplane (OSH), which is completely determined by the support vectors (right: circled dots).



knowledge about the data distribution, then the optimal separating hyperplane (that is, the one with the lowest bound on the expected generalization error) is the one which maximizes the margin [Cristianini and Taylor, 2000, Schölkopf and Smola, 2002, Vapnik, 1998]. Figure 2.3 illustrates this for two cases: in the left part of the figure, the margin - which is defined as the distance $\frac{2}{\|\vec{w}\|^2}$ of the separating hyperplane to the nearest vectors of either class - is small. In this case, small deviations from that separating hyperplane would already result in changing classification outputs and thus in a higher generalization error. If, however, the margin is made larger as in the right part of Figure 2.3, classification of test vectors becomes less sensitive to the distance to the hyperplane, which results in a smaller generalization error.

The optimal values for \vec{w} and b can be found by solving the following constrained minimization problem (maximizing the margin is of course equivalent to minimizing its inverse):

$$\begin{aligned} & \underset{\vec{w}, b}{\text{minimize}} && \frac{1}{2} \|\vec{w}\|^2 \\ & \text{subject to} && y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1, \forall i = 1, \dots, m \end{aligned}$$

Solving this optimization problem using Lagrange multipliers $\alpha_i (i = 1, \dots, m)$ results in a classification function

$$f(\vec{x}) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i \vec{w} \cdot \vec{x} + b \right).$$

where α_i and b are found by using an SVM learning algorithm (see also Cristianini and Taylor [2000], Schölkopf and Smola [2002], Vapnik [1998]). Most of the α_i 's take the value of zero; those data points \vec{x}_i with nonzero α_i are the so-called "support vectors". These support vectors define the optimal separating hyperplane (see right part of Figure 2.3) which separates the two classes and determines the decision function.¹

To obtain a *nonlinear* classifier, the so-called "kernel trick" is employed, which maps the data from the input space \mathbb{R}^N to a high dimensional feature space \mathcal{H} by $x \rightarrow \Phi(x) \in \mathcal{H}$, such that

¹In cases where the two classes are non-separable, the solution is identical to the separable case with a modification of the Lagrange multipliers to $0 \leq \alpha_i \leq C, i = 1, \dots, m$, where C is the penalty for the misclassification.

the mapped data points of the two classes are linearly separable in the feature space. Assuming there exists a kernel function K such that $K(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})$ (requirements for kernel functions are specified by Mercer's theorem and include that the kernel function is symmetric and positive semi-definite), then a nonlinear SVM can be constructed by replacing the inner product $\mathbf{x} \cdot \mathbf{y}$ in the linear SVM by the kernel function $K(\mathbf{x}, \mathbf{y})$

$$f(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \right)$$

This corresponds to constructing an optimal separating hyperplane in the *feature* space.

Several recognition experiments using SVMs will be carried out in chapters 6 and 7.

2.3 Conclusion

Interestingly, the development of computational approaches to recognition is closely intertwined with the development in the cognitive sciences, most notably in psychophysics. Whereas earlier approaches were concerned with 3D reconstruction as the basis of recognition, it was soon realized that successful vision systems could be built this way only for rather limited environments under heavy constraints. The main problem that computer vision encountered was the inherent brittleness of edge maps that constituted the basis of many 3D reconstruction algorithms (see also previous section on Marr's and Lowe's highly influential models of visual processing). Reliable extraction of robust features from real-world images proved to be extremely difficult either due to highly ambiguous interpretations of whether a feature was caused by a 3D structure or by a spurious image property. Imposing strict constraints on the extraction of features, however, severely limited the applicability in unrestricted real-world environments. In the early nineties - together with the developments towards view-based approaches in psychophysics - several appearance-based recognition systems were proposed that were able to operate on (a restricted set of) real-world images. This line of research has led to a number of highly successful computer vision approaches, which are able to recognize objects in a variety of viewing conditions. In particular, local features have emerged as a powerful and robust method for recognition with several recent approaches tackling the difficult issue of categorization (see also chapter 8).

The development on the feature extraction side was paralleled by a development in classification methods, where I specifically want to highlight Support Vector Machines as a recent example of a well-founded classification framework that offers excellent generalization capabilities.

While some of the proposed recognition systems achieved impressive recognition performance, there are two main issues, where further development can lead to increased robustness and generalization capabilities for object recognition. The first issue concerns the *combination* of the two recent developments in feature extraction and classification - local features and Support Vector Machines. The combination of these two approaches requires the development of new types of kernel functions as the standard kernel framework cannot handle local features. Chapters 6 and 7 as one of the two main computational contributions of this thesis proposes a framework for local kernels for Support Vector Machines and presents extensive recognition experiments to demonstrate the excellent recognition performance that can be achieved.

The second important issue concerns *temporal* properties of the visual input, which play a large role both during learning and recognition of objects as was discussed in chapter 1. Whereas some neural network studies have addressed this question (see for example, Massad et al. [1998]), systematic studies in the context of computer vision have been rare. In most cases, a collection of static images taken from various sources was used to represent the learned object database. In the following, I will develop a recognition framework, which actively uses the temporal dimensions both during learning and recognition. This framework will be based on cognitive research and will

be used both for modeling of psychophysical experiments as well as for computational recognition experiments.

Chapter 3

A generic framework for object learning and recognition

In the following, I want to extend and incorporate key issues from the previous chapters in order to discuss how one might envisage a computational recognition system, which would have the capability to perform cognitive tasks with a level of expertise and robustness comparable to humans. This discussion will mainly focus on issues that are still largely lacking in current recognition frameworks particularly in computer vision and that represent major open challenges for future research in this field. From this sketch, I will then extract a few key concepts in the context of the issues discussed in the two previous chapters and integrate them into a recognition framework. This framework represents one of the central contributions of this thesis. Some of the concepts presented here and developed further in subsequent chapters will also be taken up in the final chapter, where I will extend the framework towards *categorization* as well as recognition in a perception-action loop.

3.1 Introduction

First, note that within the presented framework of feature extraction and classification in the previous chapter the object model is regarded as fixed - there is no possibility for incremental learning (that is, successive improvements of the model based on additional information available in the recognition stage) or any kind of error-correction due to invalid assumptions in the learning stage (such as wrongly assigned class-labels or insufficient information for the extraction of features). In order to achieve true robustness and generalization capabilities, however, adaption and learning through a constant validation of internal models are crucial prerequisites for a successful recognition system. This essentially amounts to *closing the loop* between the learning and recognition stage. This in turn requires an additional instance that is able to judge the success of the recognition stage and based on this judgment can initiate appropriate learning processes. As a consequence, recognition should no longer be a simple yes-no decision, the system should rather adapt its internal model whenever an object is wrongly classified as recognized (false positives) or whenever an object is wrongly rejected as not recognized (false negatives) - again a decision, which has to be triggered in a supervised manner. Such a tight coupling of learning and recognition is of course also an important issue in cognitive research, where it was realized that a flexible system can only be achieved through an extensive amount of learning and feedback.

A second concept, which is important in the development of a generic computational recognition framework, is given by *structured processing* - making *structures* in the (visual) data explicit both during recognition and learning. One important element of structured processing is categorization, which establishes a hierarchy for the entire recognition process. In the context of the first chapter, categorical structure consists of scene processing, followed by super-ordinate and

basic level categorization down to recognition of single exemplars. As the search space for unconstrained object recognition is prohibitively large, a categorization hierarchy provides the necessary mechanism by restriction of the search space to a smaller number of object properties (such as all objects that share a similar appearance in the case of basic level categorization). It is interesting to note in this respect that until recently the emphasis of appearance-based computer vision research was solely on recognition, whereas earlier structured shape models were (sometimes implicitly) dealing with categorization. As was noted by Edelman and Intrator [2001], one of the drawbacks of appearance-based models is their lack of structure, which makes them unsuited for more complex spatial reasoning processes (such as determining the configuration of parts of an object or the layout of objects in a scene). This means that not only categorization but also recognition approaches should take into account structured processing. An additional element of structured processing is the introduction of context, which can help reduce ambiguities during recognition - for example, in the form of a prior on expected occurrences of objects, which belong to this given context. In the studies conducted by Torralba and Oliva [2003], participants were able to clearly identify image areas consisting of only 6 pixels as human faces if they were presented in the context of the whole human body - similarly for image areas containing cars, if they were presented in the context of a street scene. This scene context thus provides important structured cues for its interpretation. In essence, context provides a super-structure for the model database that in neural network terms could be simply thought of several excitatory and inhibitory connections between particular object models (see also chapter 8). The concept of context is, however, not restricted to the level of the whole object but can easily be extended to the processes of, for example, feature extraction or learning.

A third important concept for the generic recognition framework is given by *action or embodiment*, which implies that the recognition system is not only equipped with a number of passive input sensors but that it can also interact with the environment. As a consequence, the recognition system is able to receive direct feedback based on the current perceptual state and can thus enter a “perception-action loop”. Examples include active gaze control for robots in order to acquire additional information about particular regions of interest in the environment, or manipulation of objects in order to determine object information from other modalities (proprioception, haptics). The concept of embodied systems will be mainly considered in chapter 8.

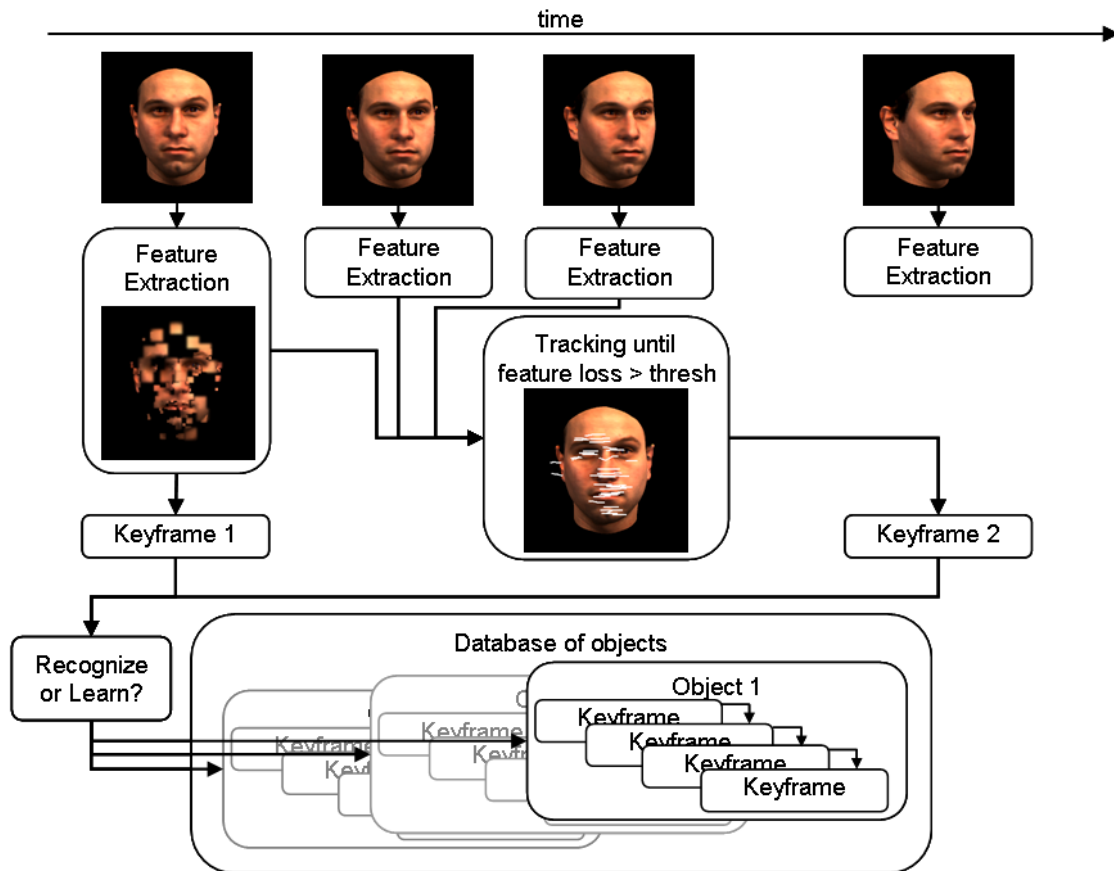
These considerations already show that in order to achieve the final goal of a truly integrated and flexible recognition system, there is still some work left to be done. Within these broad concepts I will now place several of the cognitive and computational issues mentioned in chapters 1 and 2. The goal here is to lay the foundations for an abstract object recognition framework firmly based on cognitive research, which will form the basis of this thesis and which represents a first tentative step towards a cognitive recognition system.

The first concept of closed learning and recognition requires a *robust and discriminative learning* scheme. This learning scheme should be able to generalize already from a small amount of data in order to limit training time and to start early to adapt its knowledge database to environmental conditions ¹. A closed loop in addition implies the inclusion of the *temporal dimension* for the recognition framework. Furthermore, it requires a system architecture that is capable of supporting recognition *and* learning processes thus being able to operate in an *online and iterative* fashion.

With regard to the second concept of structured processing, both architectural and procedural implications can be derived based on information from the previous chapters. First of all, we have seen that *holistic and configural* processes seem to play a large role for structuring object representations during learning and recognition. This in turn implies a development of local feature frameworks, which can support such representations (see also chapter 2). Second, as was

¹This description share much in common with the so-called “tabula rasa” concept, i.e., the massive reorganization of neural connection in the early developmental stage of children, which has been shown to be highly dependent on external stimulation through sensual input.

Figure 3.1: Abstract description of the keyframe framework.



discussed in chapter 1, *spatio-temporal continuity* provides an important structuring principle for acquiring view-based representations from dynamic visual input. A third, more general, structuring element is given by the *spatio-temporal processing* of objects, which requires both appropriate local features as well as classification algorithms.²

In the following sections, I will describe a computational recognition system that implements these key concepts outlined above. In this chapter I will mainly focus on the structural properties of the system which - independent of a specific computational implementation - already will prove valuable in the cognitive modeling experiments described in the next chapter. Chapter 5 will then provide more computational experiments in order to quantify the performance of the system both with real-world and controlled databases.

3.2 Learning and recognizing objects using keyframes

The abstract framework shown in Figure 3.1 implements several key elements from the previous section, which I will explain in the following.

On an abstract level, the system processes incoming frames from an input sequence in a sequential manner in order to extract so-called *keyframes*. Keyframes are views - snapshots - of the scene, which are defined by the *temporal continuity* of the visual input and represent an extension of the view-concept that lies at the basis of the view-based approach [Tarr and Bülthoff,

²As mentioned, *categorical hierarchies* also provide structure on an architectural level making robust recognition of a large number of objects feasible. The proposed framework will focus mainly on exemplar recognition with some extensions towards enabling categorization of objects presented in the outlook of this thesis.

1998]. In essence, keyframes divide the sequence into temporally continuous "chunks", where it will be of course important to specify exactly which properties of the visual input can define and support temporal continuity *per se*. Equivalently, keyframes can be understood as points in the image sequence, where a visual event occurred. The inclusion of the temporal dimension is motivated by the results on temporal association by Wallis and Bühlhoff [2001], Wallis [2002] discussed in chapter 1. The main idea behind this framework is that the sequence is completely specified by the temporal sequence of keyframes, which thus form *a connected graph of views* (see Figure 3.1).

More specifically, In the first frame of an image sequence, *local features* are extracted at several scales, which are then *tracked* in the subsequent frames. Once tracking fails, it is assumed that a visual event has occurred. At that point in time, a new keyframe is added to the representation and a new set of features is extracted in this keyframe, and the whole process repeats until the sequence ends. The final representation of the sequence then consists of a number of keyframes containing visual features on multiple scales. For recognition of test images as well as for checking whether to add keyframes to existing representations, the local features and their configurations are matched against all keyframes in previously learned representations using robust classification algorithms. In the proposed framework, learning and recognition are therefore in principle not separated, which means that the system constantly learns and acquires new data, which in turn can be used to augment existing object representations or to form new ones.

3.2.1 Related concepts

Note that in this abstract form the keyframe approach bears resemblance to two other concepts in the computer vision literature: the first is the "aspect graph" framework by Koenderink and van Doorn [1979], in which objects were defined by their aspects, that is, by visual events, where a sudden change in the observed shape of the object occurred. The original work used well-defined geometrical properties of the observed objects (such as curvatures and segments) that caused such visual events. The aspects of an object could thus be used to determine its shape under certain geometric constraints. Even though the mathematical formulations were highly appealing to the computer vision community due to their geometric interpretations, computational realizations of the aspect framework for *arbitrary* objects proved to be difficult. There has been relatively little work on how to extend this concept into a robust recognition framework, most notably by Cyr and Kimia [2001] who developed a 2D version of the aspect graph framework based on shape similarity. The proposed keyframe framework represents an important extension of this work as it includes appearance-based processing of local features in a spatio-temporal context.

Furthermore, I want to go one step further with the keyframe concept by representing *all kinds* of dynamic visual input with the help of 2D views. Although this of course also includes the standard case of viewing sphere exploration (see for example, Peters et al. [2002] for a recent study as well as chapter 5), I want to adopt a more general approach, which represents more generic types of visual input as a sequence of connected views. Examples mentioned already in chapter 1, which go beyond the usual 3D rotation paradigm, include *dynamic* changes in the sequence such as illumination changes, or object deformations. There are relatively few studies that have directly used spatio-temporal information for recognition in computer vision - most notably the study by Li and Chellappa [2001] who have employed tracking of a fixed grid of points to gather evidence for recognition of faces - and I would like to provide the keyframe representations as one potential candidate for such a representation given the psychophysical motivations outlined in the previous chapters. In addition, I want to mention the fact that the proposed concept of keyframes is very well suited for multi-modal representations combining different modalities such as haptics and vision and thus brings the *active* component of visual processing to bear (see chapter 8).

The second related line of research I want to mention in the context of keyframes, consists of image sequence encoding and decoding through lossy compression techniques such as

MPEG1,2,4 (see Richardson [2003]), which have attracted some attention recently. Here the goal is to compress streaming image data as much as possible while still retaining good visual quality. Several of the techniques available also use a "keyframe" approach, whereby the complete sequence is represented as a number of (evenly spaced) keyframes and motion vectors, which specify how image elements move from one keyframe to the next. It is exactly this concept, which inspired the name for our proposed framework, although in this case the aim is not to reconstruct image sequences but rather to *recognize spatio-temporal patterns* through sparse information³.

3.2.2 What defines a keyframe?

In essence, the question of how to find properties of the visual input, which can be used to support temporal continuity and to define keyframes, is reduced in the proposed framework to *local feature processing* (see Figure 3.1). This includes methods for extraction of salient visual features as data representations and in addition algorithms for processing these features, all of which can support learning and recognition of spatio-temporal visual input. The basis for extracting keyframes is to process each input frame of an image sequence in order to extract *interest points* and to use these interest points in a *tracking framework* to determine the segmentation of a sequence into keyframes. Furthermore, as pointed out in the introduction, these visual features will also be used in a *structured, local feature recognition* framework, which supports both configural and component recognition processes (see chapter 1) thus increasing the robustness and discriminability of the object representation.

An interesting question in this context is, whether there is evidence from cognitive studies for local feature analysis beyond the component processes outlined in chapter 1. Whereas there exists a wide range of interest point detectors in the computer vision literature (see chapter 2), very few papers directly address psychophysical or physiological validations of these detectors. One standard pre-processing step, which seems to be biologically plausible is to filter the input image by banks of oriented filters (usually Gabor filters), which are modeled after properties of the receptive fields of early visual areas (see section 1.3 and also Riesenhuber and Poggio [1999] for a recognition framework that is based on these kinds of features). In addition, physiology studies support the notion of *features of intermediate complexity* [Ullman et al., 2002], that is, compound features made up of smaller, simpler features. More direct evidence about the nature of such features comes from studies by Krieger et al. [2000] who analyzed foveation locations on natural images (see also an earlier study by Kaufman and Richards [1969]) and found that "inherently 2D image structures" could be used to model these locations. In essence, this studies support the idea that during visual processing, our visual attention is drawn to "interesting" points in an image, which are characterized by a high 2D intensity contrast.

3.3 Discussion of the framework

Before going into implementation details, I want to analyze the proposed framework in more detail in order to answer one of the key questions: What - apart from the psychophysical motivations - is the advantage of using the temporal dimension in the framework? Following the argumentation of Wiegardt and von der Malsburg [2000], learning and recognition certainly seems possible using only the *static* dimension. Nevertheless, there are several important aspects of this framework, which can be expected to lead to increased performance concerning both efficiency and accuracy of matching.

³In principle, one could also argue that image sequence compression techniques might enable "recognition by synthesis" Blanz and Vetter [1999] - although this would be an interesting research area, the main goal of this thesis is to introduce an abstract framework motivated by and realizable in cognitive systems.

3.3.1 Keyframes

In the most extreme case of a view-based framework, learning would involve storing *all* input images (see chapter 2). This strategy does not seem feasible for two main reasons: first, unconstrained object recognition in a view-based framework requires large amounts of learning data as different kinds of object changes (lighting, pose, etc.) will need to be encoded using views. Storing all input images would therefore result in an exponential growth of the database and would quickly exceed even generous storage capacities. Second, a large database of views represents also a severe problem for recognition, as the time it takes to *index* into the representation in order to retrieve views for recognition becomes an issue.

The important question thus is: which views to select for learning? Here the keyframe concept provides an intuitive answer to that question: select views, where an important visual *event* occurs. In order for this strategy to be successful, one needs to make the assumption that the visual input is on average slowly changing - or equivalently that processing of visual input is fast enough for this assumption to hold. Given the psychophysical evidence by Wallis and Bülthoff [2001] presented in chapter 1, this certainly seems to be a valid assumption. During this selection stage the system could operate in a "simple" mode, which only uses information from the immediate visual past, as each incoming image is evaluated based on the current keyframe. This also means that the system does not have to access its full database of stored object representations during learning and thus uses only little (computational) resources.

This assumes, however, that once a visual event occurs and a keyframe is generated during learning, it is simply inserted into the database of stored representations with a link to the previous keyframe. This strategy of course has a major disadvantage: letting the system run for only a short amount of time will produce a huge amount of data as all generated keyframes have to be stored. The crucial point here is that there is no test whether generated keyframes have already been learned in the recent past (for example because an object passes through the same view twice) or whether they might already be part of an older, different representation in the database (for example, by seeing an object again after some time). As far as the first point is concerned, a simple strategy is to try and match each newly generated keyframe against all keyframes in the existing object representation - that is, to implement an incremental learning strategy (see chapter 5). With regard to the second point, one has to make the distinction between supervised and unsupervised settings: in a supervised learning setting, labels are available for learning and thus one could simply resort to the incremental learning strategy mentioned above. In an unsupervised setting, however, one would have to match such newly generated keyframes against all previously learned keyframes. This problem would also occur in the recognition phase of the system: each image (or keyframe) that has to be recognized requires a full search through the database of learned representations, which again highlights the problem of recognition in large databases.

The structure of the keyframe framework, however, provides a solution for this case: as shown in Figure 3.1, keyframes are organized in a directed graph structure. Considering the case in which two connected keyframes in a row could be recognized, chances are good that the next incoming keyframe will be the next node in this graph. Thus, if a keyframe was already recognized, it would make sense to first match any newly generated keyframe or image to the outgoing links of that keyframe. This strategy thus dramatically reduces the search time during recognition of known sequences or sequence-parts. There is still a possibility, however, that a frame could *not* be recognized - in this case there are two possibilities:

- it is simply inserted either as a starting point of a new object representation
- in a supervised learning setting - which means that labels of images are available - it is attached to an existing object representation without a link to the remaining keyframes in that representation.

This incremental learning strategy thus results in a continuously learning system, which gathers

visual data in the form of linked keyframe representations (which in the supervised learning case correspond to object representations). The assumption behind such a "curious" system is that the accumulated visual knowledge at some point will be saturated with a stable set of representations.

An interesting question in that case would be, what such a saturated representation would look like. One possible answer might be provided by the studies on canonical views reviewed in chapter 1: if, indeed, there exist views of objects which are canonical, this might be due to either their visual saliency (that is, these views are well suited for being recognized) or due to visual frequency (that is, they are simply often seen). Both of these properties would result in nodes (keyframes) in the object graphs that have a large number of links and that in addition were often matched - a network of keyframes. Whereas it would certainly be interesting to conduct such an extensive study with a system that would learn keyframes for days and even years, this is outside the scope of this thesis (see chapter 5 for smaller-scale studies).

Going a few steps back, however, in the next section I want to discuss the bottom-up information, which enables the generation of keyframes - local visual features.

3.3.2 Local visual features

The choice to include local visual features in the framework first of all reduces the size of the representation while at the same time increasing the discriminability of the input representation for recognition (see chapter 2 for a discussion). Given the large body of research on local feature types, a broad range of features can be integrated with our approach - from simple image fragments [Ullman et al., 2002] which still preserve some of the visual appearance of the object to more abstract histograms of local filter responses [Lowe, 2004, Mikolajczyk and Schmid, 2005, Schmid and Mohr, 1997, Schiele and Crowley, 2000]. Two important requirements that have to be met in order for a local feature approach to be compatible with our framework are: first, a number of interest points have to be extracted in an image and, second, some sort of local descriptor is computed around each of these points. These steps are necessary as the extracted local features have to support *tracking* as well as *recognition*.

The most important contribution of *feature tracking*, which is used to determine keyframes, is that it allows access to the feature *trajectories*. In our framework, such trajectories trace features from one keyframe to the next. The larger the visual difference between keyframes is, the more discriminative these feature trajectories are, as the chances of false matches is reduced the longer a feature can be reliably tracked (see also Tomasi and Kanade [1991]). In addition, the trajectories describe the transformation of each feature from one keyframe to another and thus can be used to generate *priors* for matching local feature sets. In order to illustrate this, let us consider a sequence of a rotating object. The image trajectories for the keyframes of that object will have a very uniform shape, which is specified by the direction of the (3D) rotation. For recognition, a *matching prior* can now be derived directly from the trajectories by constraining feature matches to that very direction.

In addition to enabling access to the temporal dimension, local features allow the framework to support configural and component processing, which was shown to be an important recognition approach for (face) recognition in chapter 1. As the class of visual transformations that can occur in a given viewing situation (see Figure 1.1, 1.2) is quite large, additional constraints on the structural organization of visual elements in the image can help both to disambiguate matching processes and to provide increased and more robust matching performance across a larger range of visual transformations (such as rotations in depth, etc.). Again, it is important to keep in mind that - given the evidence from cognitive studies - achieving *invariance* to such transformations is not the goal of this framework. Furthermore,

In addition, I want to stress that this focus on visual features and their transformations between visual events is a much broader concept not restricted to object motion alone. Going beyond a simple matching prior, information about spatio-temporal changes in appearance could also be

used to explicitly model *category transformations* (see Graf et al. [2002] and chapter 8) or object transformations due to *illumination* changes, which transforms the keyframe framework into a potentially generic learning concept for *any* kind of dynamic visual data.

3.4 Computational implementation

In the following I will provide a more detailed description of a possible computational implementation of the keyframe framework. Again, the focus of this chapter does not so much lie on developing novel algorithms for feature extraction and/or tracking but rather on describing how the general architecture shown in Figure 3.1 can be efficiently implemented.

3.4.1 Visual features

In computer vision, large scale studies done by Mikolajczyk and Schmid [2005], Schmid et al. [2000] on stability of interest point detectors to noise, scale, illumination and image plane warping found that the best performance could be obtained with a modified standard Harris [Harris and Stephens, 1988] corner detector. In the following computational experiments a similar type of corner detector was used in the context of a *multiple scale framework* as a more abstract approximation of what the human visual system might work with - the keyframe framework proposed above, however, can of course support all kinds of different features.

One possible implementation of such an interest point detector is based on a well-known algorithm proposed for tracking features [Tomasi and Kanade, 1991]. This algorithm is extended to process input images in a multi-scale framework using a Gaussian scale pyramid [Duda et al., 2001]. In order to extract interest points for a given sub-image of the scale pyramid, the following matrix \mathbf{H} is evaluated at each point in the image:

$$\mathbf{H} = \begin{pmatrix} \sum_{\mathcal{N}} \left\langle \frac{\partial \vec{I}}{\partial x}, \frac{\partial \vec{I}}{\partial x} \right\rangle & \sum_{\mathcal{N}} \left\langle \frac{\partial \vec{I}}{\partial x}, \frac{\partial \vec{I}}{\partial y} \right\rangle \\ \sum_{\mathcal{N}} \left\langle \frac{\partial \vec{I}}{\partial x}, \frac{\partial \vec{I}}{\partial y} \right\rangle & \sum_{\mathcal{N}} \left\langle \frac{\partial \vec{I}}{\partial y}, \frac{\partial \vec{I}}{\partial y} \right\rangle \end{pmatrix}$$

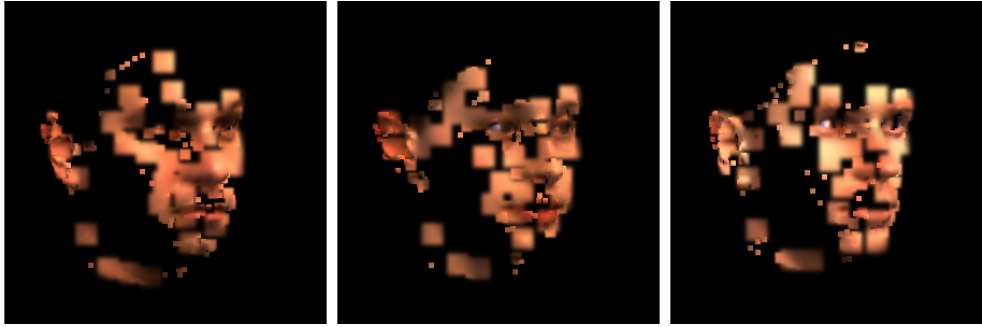
with \langle, \rangle as dot-product and \vec{I} as the vector of RGB-values such that an element of \mathbf{H} is for example $H(1,2) = \sum \frac{\partial^2 I_r}{\partial x \partial y} + \frac{\partial^2 I_g}{\partial x \partial y} + \frac{\partial^2 I_b}{\partial x \partial y}$. The smaller of the two eigenvalues λ_2 of \mathbf{H} yields information about the structure of the neighborhood. Thresholding λ_2 results in a set of positions \vec{p}_i in the image for which a significant change in intensities could be found.

In a local feature framework, the basic idea is to combine a set of salient interest points with a suitable feature descriptor, which characterizes the visual appearance of the image around these points. One possibility to do this, for example, is to simply take the *pixel neighborhood* \mathcal{N} around each interest point as the feature descriptor - one could also call this a fragment-based approach (see Ullman et al. [2002]). As the extraction of visual features is done at several resolution scales, this process results in visual features with small details at fine scale levels and with large details at coarse scale levels.

The final representation of each image then consists of n interest points (represented by their pixel coordinates) and their surroundings (represented by small image fragments). Such a representation is shown in Figure 3.2, where each of the three faces was reconstructed from its feature representation by successively adding the image fragments at each scale - starting from a coarse level with large image regions and progressing to the finest level with small, detailed features. Note also, that already this simple interest point detection scheme tends to focus on important facial features such as eyes, mouth and ears.

This representation bears resemblance to lossy image compression techniques such as wavelet compression [Mallat, 1989] but with added spatial information about the location of points with high contrast information. Indeed, the compression rate of the images is high with respect to the original

Figure 3.2: Reconstruction of views of three faces from their visual feature representations.



file-size considering that the image content still allows for visual reconstruction. With an average number of visual features of around 200 over all resolution levels as well as a neighborhood-size of 5x5 pixels, the resulting size of the representation is 14.6 Kilobyte for a 256x256 pixel RGB image of originally 192 Kilobyte (both figures indicate raw data size with data stored in an uncompressed format).

3.4.2 Matching of visual features

Since the proposed framework is based on visual features, which should form both the basis for *recognition and learning*, an algorithm for matching sets of visual features is needed. Whereas in principle one could have chosen different algorithms for the extraction of keyframes (that is, tracking) and for recognition, here I want to introduce a single algorithm that can support both tasks. This leads to an efficient implementation, which further blurs the distinction between the learning and the recognition steps.

3.4.2.1 Constructing a similarity matrix

The method, which we shortly summarize, is based on work by Pilu [1997] (where it was used for stereo matching) and Scott and Longuet-Higgins [1991] and is a variation on the well-known Procrustes problem of rotating two datasets onto each other.

First, a similarity matrix \mathbf{A} is constructed with each entry $A(i, j)$ given by two contributing terms:

$$A(i, j) = \delta_{\text{geo}}(\vec{p}_i, \vec{p}_j) \cdot \delta_{\text{app}}(\vec{f}_i, \vec{f}_j) = \exp\left(-\frac{1}{2\sigma_{\text{geo}}^2} \text{geo}(\vec{p}_i, \vec{p}_j)\right) \cdot \exp\left(-\frac{1}{2\sigma_{\text{app}}^2} \text{app}(\vec{f}_i, \vec{f}_j)\right)$$

where δ is a suitable similarity metric between two vectors, \vec{p}_i is the position and \vec{f}_i the local descriptor of feature $\vec{f}_i = (\vec{p}_i, \vec{f}_i)$ in one image and \vec{p}_j is the position and \vec{f}_j the local descriptor of feature $\vec{f}_j = (\vec{p}_j, \vec{f}_j)$ in the second image with i, j indexing all feature pairs in the two images.

The first term measures an image-based, position similarity (geo), which depends on the 2D geometric relationships between feature i and feature j , whereas the second term measures an appearance-based similarity (app) between the two features, which only depends on the local descriptors of each feature. In the simplest case, the local feature descriptors simply consist of the image patches transformed to a vector of image intensities - other, more sophisticated descriptors are of course also possible. Both terms are cast into an exponential form, which yields a positive definite matrix \mathbf{A} and which further allows to specify a typical scale for the similarity measures through the standard deviations $(\sigma_{\{\text{geo}, \text{app}\}})^4$.

⁴In principle it is sufficient to specify only the *relative* weighting of one similarity metric against the other thus reducing

3.4.2.2 Finding corresponding features

The next step is to find corresponding features between images I_i and I_j given the similarity matrix \mathbf{A} . Whereas it might be desirable in some cases to allow a one-to-many matching of features, for the purpose of recognition a constraint on one-to-one matching seems reasonable in order to reduce the possibility of false matches. The simplest way to achieve this is to find the largest elements both in row and column in \mathbf{A} , where the number of matches is limited to $n = \min(n_i, n_j)$. This “greedy” strategy has a computational complexity of order $O(n_i n_j)$ and thus is feasible in all of the recognition scenarios examined here, where $n = O(100)$.

A more elaborate and computationally more intensive correspondence algorithm is based on work by Scott and Longuet-Higgins [1991] and relies on the Singular Value Decomposition (SVD) of the matrix \mathbf{A} . From the SVD, which is defined as $\mathbf{A} = \mathbf{U} \cdot \mathbf{V} \cdot \mathbf{W}^T$ the matching algorithm constructs the modified SVD of this matrix, defined by $\mathbf{A}' = \mathbf{U} \cdot \mathbf{I} \cdot \mathbf{W}^T$, where \mathbf{I} is the identity matrix. The replacement of \mathbf{V} with the identity matrix has the effect of normalizing the Eigenspace of the matrix thereby weighting each dimension equally. Features i and j are again matched if they have both the highest entrance in the column and row of \mathbf{A}' . This method effectively provides a least-square mapping with respect to the two similarity terms and at the same time ensures that there is a one-to-one mapping of features⁵.

For both types of match schemes, it is possible to introduce an additional match threshold, which is imposed in the last step: the match is accepted only if $\max_{i,j} \mathbf{A} > thresh$. This threshold will further enhance match fidelity and provide additional robustness - which of course comes at the price of introducing an additional parameter. Furthermore it is possible to impose the final match threshold not on the combined similarity value, but directly on either $\delta_{geo}(thresh_{geo})$ or $\delta_{app}(thresh_{app})$. Specifying such a threshold for one similarity value only might at first glance seem to make the matching scheme more “asymmetric”. Justification for this prioritization, however, can especially be given for a $thresh_{app}$ as favoring the appearance of a feature fits well with the concept of appearance-based or image-based matching (see Pilu [1997]).

For recognition and tracking purposes, the number of matches is counted in a test frame with respect to the given reference frame and the *percentage of matches* is used as the decision criterion for matching or tracking. The resulting algorithm was shown to be capable of matching under affine transformations in the *similarity measure* space (see Scott and Longuet-Higgins [1991], Pilu [1997], Wallraven and Bühlhoff [2001a,b]), that is, when using $\delta_{geo}(\vec{f}_i, \vec{f}_j)$ only, it would be able to match features which undergo an affine transformation in pixel coordinates. Note also, that in principle the algorithm does not make any *explicit* assumptions about the type of transformation occurring between two feature sets. This makes the algorithm ideally suited for the modeling and computational experiments of this thesis as it does not impose severe restrictions on the possible types of similarity measures.

3.4.2.3 Examples of similarity measures

In the following I will introduce the most basic versions of the two similarity terms as they will be used throughout the experiments.

The simplest form of the first term of the similarity matrix, which measures 2D configurational information of features, is given by the Euclidean pixel distance:

$$geo(\vec{p}_i, \vec{p}_j) = \|\vec{p}_i - \vec{p}_j\|^2$$

This results in a matching bias towards close matches in the 2D image plane. It is, however, possible to introduce other more psychophysically plausible configurational terms, which will be outlined in the following chapter.

the number of parameters to only one.

⁵Note that the feature mapping can occur between sub-sets of features.

Similarly, the most simple form of appearance similarity is given by the standard Euclidean distance between the two feature vectors:

$$\text{app}(\vec{f}_i, \vec{f}_j) = (\vec{f}_i - \vec{f}_j) \cdot (\vec{f}_i - \vec{f}_j)^T$$

For the appearance term, it is also possible to introduce a slightly more complex similarity measure, which is more suitable for working with image patches as local descriptors. Setting $\vec{f} = \mathcal{N}_{\vec{p}}(I)$ as the square pixel region centred at position \vec{p} in image I , this appearance-based similarity measure is the normalized cross correlation defined as:

$$\text{app}(\vec{f}_i, \vec{f}_j) = \text{NCC}(\vec{f}_i, \vec{f}_j) = \frac{\sum_{i \in \mathcal{N}_i, j \in \mathcal{N}_j} (I_i - I_{m,i}) \cdot (I_j - I_{m,j})}{\sum_{i \in \mathcal{N}_i} (I_i - I_{m,i}) \cdot (I_i - I_{m,i}) \sum_{j \in \mathcal{N}_j} (I_j - I_{m,j}) \cdot (I_j - I_{m,j})}$$

where $I_{i,j}$ are intensity values, $I_{m,i}, I_{m,j}$ mean intensity values in square image regions $\mathcal{N}_i, \mathcal{N}_j$ centered at \vec{p}_i, \vec{p}_j respectively. The NCC yields values between -1 (completely dissimilar) and +1 (identical) for each feature. It also follows from the definition that the NCC is invariant under linear transformations

$$\mathcal{N}_1 \rightarrow a \cdot \mathcal{N}_1 + b, \mathcal{N}_2 \rightarrow c \cdot \mathcal{N}_2 + d \quad a, b, c, d \in \mathcal{R}$$

of image intensities *within* $\mathcal{N}_{1,2}$. This property could for example provide increased stability under lighting variations since globally *nonlinear* intensity changes due to lighting variations can be approximated as *linear* within \mathcal{N} . Using local features thus makes this framework more flexible than appearance-based approaches working with whole images.

3.4.3 Recognition and Incremental Learning

As mentioned earlier, the proposed framework in principle allows to integrate the learning and recognition stage. Based on the discussion in chapter 2, one can identify a number of possible modes in which the framework could be run: Learning can either be done in a *fixed* context in which the number of sequences is pre-determined or in an *incremental* (or online) context in which representations of objects can be extended or new objects can be learned. In addition, recognition as well as learning can be done in a *supervised* manner where class/object labels are known or in an *unsupervised* manner. The ultimate goal, of course is to have a fully unsupervised system that learns and recognizes objects online.

The fixed/supervised mode corresponds to the standard setup for most computer vision experiments. Here, a database of labelled sequences is first encoded into keyframe representations. In the case of supervised recognition, another database of labelled test sequences is encoded into keyframes, which are then compared to the stored representations using any of the classification algorithms discussed in chapter 2. Similarly, unsupervised recognition would amount to finding labels in the learned database for each of the encoded keyframes - analysis of recognition performance would, of course, not be possible in this case.

For incremental/supervised learning, the procedure would be very similar as each incoming keyframe possesses a label which allows clear decisions about class membership. The advantage of this mode - apart from being able to extend existing representations as well as learn new ones - would be that already learned keyframes for known objects would not be duplicated. Note, however, that for this each new keyframe *during learning* requires to be checked against all previously learned keyframes that carry the same label, that is, that belong to the same image sequence. Learning and recognition stage are therefore not fully separable in this case. In addition, if classification is done using more complex classifiers such as Support Vector Machines, incremental learning requires to re-train (or update) each time the representation is extended or a new representation is inserted.

As soon as learning takes place in an unsupervised setting, no clear distinction between learning and recognition is possible, as each new keyframe needs to be checked against *all* previously learned keyframes in order to decide whether to start a new representation or not.

Interestingly, the temporal context of keyframe learning allows one to characterize the performance of the different modes by only two thresholds: a threshold $thresh_k$ that specifies how many features can get lost until a visual event is triggered, and a threshold $thresh_r$ that specifies when to reject a new keyframe from a test sequence as *not* belonging to the learned keyframes⁶.

3.5 Conclusion

In this chapter a framework for learning and recognition of view-based, spatio-temporal object representations was presented. The key concepts used in development of this framework are:

- close coupling of object learning and recognition
- sparse visual representations through local features
- tracking of local features for explicit access to feature transformations
- feature matching based on local appearance and layout information
- "online" approach - extensible representations

These concepts were derived from and motivated by desirable properties of cognitive vision systems for object recognition, which should enable robust learning and recognition for cognitive systems. In the following chapters several instantiations of this framework will be used to model several cognitive studies as well as to demonstrate the advantages of these cognitive concepts in computer vision and machine learning contexts.

⁶Note that the discussion in this section assumes that *all* visual input will be learned by the system. If this is not desirable, an additional threshold is required, which specifies when to start a new representation.

Chapter 4

Cognitive modeling studies

In this chapter, the recognition framework outlined in the previous chapter will be used to model human performance characteristics found in psychophysical experiments on face and object recognition as described in chapter 1. The main motivations for this chapter are

- to validate the assumptions on which the framework is built and to verify that it can be used to model the psychophysical results
- to use the computational modeling results to suggest further psychophysical experiments, which address important questions on the processes of human object recognition, and
- to use the results from the psychophysical experiments to suggest ways to improve and extend the computational system by optimizing the framework and its parameters.

The second point could be regarded as closing the loop between computational and cognitive modeling - an essential step in increasing our understanding of the complex processing behind object recognition. Closely connected with that is the third point, which should demonstrate how to use this knowledge in order to create more efficient and effective frameworks for computer vision tasks.

4.1 View-based recognition of faces

To briefly recapitulate the main findings from section 1.2.2, in a psychophysical experiment by Wallraven et al. [2002], we were interested whether view-dependency patterns found in earlier studies for unfamiliar object recognition [Bülthoff and Edelman, 1992] would also generalize to one of the most overlearned object classes: faces. The experimental design closely followed those of the original study with four testing conditions labeled as (inter, extra, ortho up, ortho down), which specified the change on the viewing axis with respect to the training view, with which novel views of a face had to be recognized.

In accordance with Bülthoff and Edelman [1992] it was found that

- recognition in the inter condition was better than in the extra condition
- recognition in inter and extra conditions was better than in both ortho conditions
- performance did not differ in the two ortho conditions.

This pattern of results is difficult to reconcile with some proposed models of object recognition (alignment of a 3D representation Lowe [1985] and linear combination approach for face recognition Ullman and Basri [1991]). However, the results can for example be understood by a linear interpolation within an RBF network [Poggio and Edelman, 1990] and as I will show now can also

be modeled within the proposed computational framework. Both of these frameworks predict $\text{inter} > \text{extra}$, which was shown clearly in the psychophysical data. Furthermore, the experimental results seemed to indicate the existence of a horizontal viewing prior which provided an explanation for why performance in the extra condition was better than performance in the ortho condition [Bülthoff and Edelman, 1992]. Interestingly, the results also demonstrate that familiarity with the object class does not necessarily predict qualitatively different viewpoint dependence, which lends additional support to the view-based account of recognition.

Going one step further, in another line of experiments, the task was extended along two dimensions. First, learning took place also on the vertical pose axis instead of the horizontal axis and second, all experiments were repeated with *inverted* faces. With the first manipulation it was possible to provide further support for the existence of the horizontal viewing prior, as performance in the different conditions did not remain $\text{inter} > \text{extra} > \text{ortho}$, but changed to $\text{inter} = \text{ortho} > \text{extra}$. This asymmetric treatment of the viewing sphere seems to be the result of a prior which, for faces, is less sensitive to horizontal changes in pose. In order to test whether this prior might be the result of low-level stimulus information (symmetry of faces) or of a higher-level learned information (that is, a learned face specific prior), the previous experiments were repeated with inverted faces. The combined results of these experiments showed a significant effect on performance ($\text{inter} > \text{extra} = \text{ortho}$) which seemed to have destroyed any large-scale influence of a horizontal prior.

The main characteristics of human view-based performance concerning faces can thus be characterized by the following elements:

- a horizontal prior that is effective for upright faces
- recognition processes based on a combination of trained views

In the following computational experiments, I want to investigate how these results can be modeled within the proposed keyframe framework. In order to do this, in general it will be first necessary to decide whether a test image is an upright face or not. As there are excellent pose detection algorithms available in the literature (see Hjelmas and Low [2001])¹, I want to focus only on the *second* element, where the test image is matched to a learned view representation taking into account the horizontal matching prior.

4.1.1 Feature matching - the horizontal prior

One of the explanations for the observed pattern of performance with regard to the difference between the *inter/extra* versus the *ortho* condition is a *learned* prior for horizontal viewing changes for objects. A possible reason for this prior might be the biomechanical constraints of human head movement, which result in a much larger freedom for azimuthal than for elevational rotations. This prior could thus be learned from the spatio-temporal statistics of human visual input of seen faces, which represents a strong link to the proposed keyframe framework, where this statistic would for example be reflected in the distribution of keyframes across the (viewer-centered) viewing sphere.

The framework proposed in chapter 3 provides a straightforward way how to implement such a prior: As discussed in section 3.4.2, in order to match two sets of visual features, a pairwise similarity matrix \mathbf{A} is constructed. The prior can then be taken into account by modifying the distance term in the matching equation such that it penalizes deviations from the horizontal direction for feature matches by an increased weighting of the *vertical* distance between features i and j :

$$A(i, j) = \exp\left(-\frac{1}{2\sigma_{\text{dist}}^2}((x_i - x_j)^2 + a(y_i - y_j)^2)\right) \cdot \exp\left(-\frac{1}{2\sigma_{\text{NCC}}^2}\text{NCC}(i, j)\right), \quad a > 1$$

¹Note, however, that it would be possible to use the configural route as presented in section 4.2 for such a pose detection task as it seems to capture the generic layout of a face fairly well.

Figure 4.1: Effect of the horizontal matching prior a on matching quality - colors indicate matches from different scales: a) $a=1$ b) $a=3$ c) $a=10$

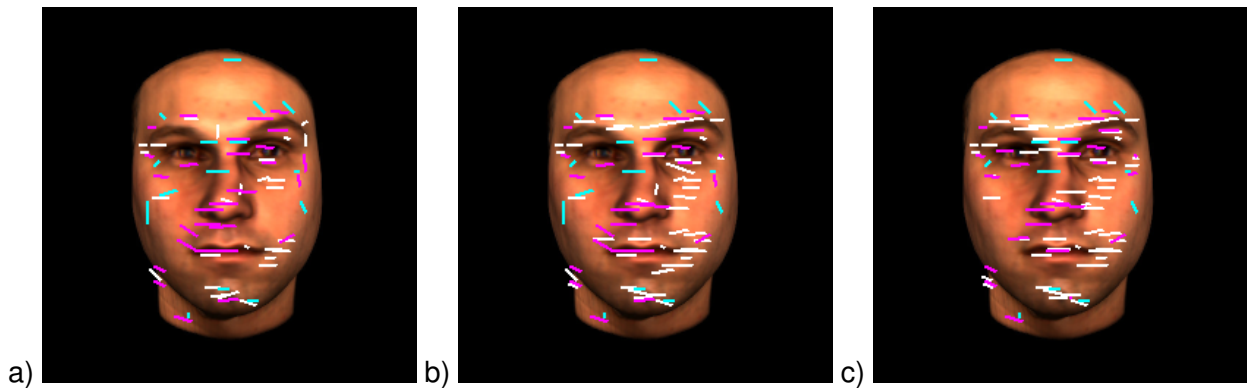


Figure 4.1 shows matching features between two images for settings of $a=1$, $a=3$, $a=10$ (Figure a)-c)): $a=1$ yields a matching score of 39 percent, $a=3$ yields a matching score of 41 percent and $a=10$ 45 percent. In the following, I will use this basic local feature matching scheme in order to test its performance on the face images used in the psychophysical experiments.

4.1.2 Modeling psychophysical experiments

Experiment Design: In order to compare machine and human results, the system was trained with two images of 10 target faces on the *horizontal* viewing axis (0° and 45°) and tested on the *same* views as humans with 10 target and 10 distractor faces. The 10 target faces were learned by extracting visual features as outlined above and subsequent storing of these representations. In the testing phase, face images were presented to the system, which again extracted the feature representation and then found the highest matching score among the 10 learned faces using the local feature matching procedure. The presented face could be either a target or a distractor and was shown in one of the thirteen views on the horizontal and vertical viewing axes (cf. Figure 1.5). In order to get a better statistical sampling, the experiment was repeated 10 times, each time with a different set of target and distractor faces. Both target and distractor faces were identical to the ones used in the psychophysical experiments.

In a next step, the experimental data was converted into a performance measure that can be directly compared to the psychophysical data. For this, the matching scores were converted into an ROC-curve by thresholding the matching scores for the target faces (resulting in hit-rates as a function of the threshold) as well as the matching scores for the distractor faces (resulting in false-alarm-rates as a function of the threshold). Finally, the area under the ROC-curve was measured, which in this case yields a non-parametric measure of recognition performance (again, $0.5 \leq \text{AUC} \leq 1.0$) that can be compared directly with the psychophysical data.

One of the most critical problems in trying to repeat human experiments with a computational system consists of the fact that the computational system can never aim at modeling all levels of human performance. First of all, humans are extremely experienced observers with a considerable expertise for recognizing faces - a fact which was not explicitly modelled in the computational implementation. Second, humans do not have perfect memory, which results in errors in recall of learned patterns compared to computational implementations. Finally, the fact that humans already learned to recognize a considerable amount of faces during their life could also result in a decrease in performance for any old-new task, since confusions of previously learned faces and experimentally learned and tested faces would be expected. In this context, the generalizability of the computational results was investigated by repeating the recognition experiment with *20 faces*

instead of 10 faces. Using more faces results in a more difficult decision for the computational system, which in turn allows a more general assessment of the influence of *learning complexity* during training and testing.

All results were compared against the baseline of a standard L2-matching that uses a *holistic* approach to image matching (the Euclidean distance between two images) rather than the proposed local feature matching strategy. Finally, the parameter a which controls the strength of the horizontal matching prior was set to $a = 0, 10, 100, 1000$ in order to gauge the influence of the prior with respect to human performance.

Results+Discussion: Figure 4.2 shows the results from the computational experiments in relation to human performance. The graphs in the left column show how performance changes as a function of the strength of the horizontal matching prior the influence for each tested view. The graphs in the right column present the averaged computational and human performance for each of the four regions on the viewing sphere.

As the left graph in Figure 4.2a) shows, the computational model performs well in all views of the inter conditions, as well as for all 15° degree views in both the extra and ortho conditions. As soon as the change in viewing angle becomes larger than 15° , however, performance drops considerably in these conditions. Increasing the influence of the horizontal matching prior helps to increase performance in the extra condition and, for a value of $a = 1000$, even allows for robust recognition of the 45° view in the extra condition. As expected, increasing a also has the consequence of decreasing recognition performance in both ortho conditions.

Turning to the right graph in Figure 4.2a), which plots the averaged results for both the computational and human data, it is first of all obvious that the computational model is not capable of reaching the same level of performance as humans in the upright condition. Whereas for large values of a , performance in the inter and extra conditions approaches human levels, performance in both ortho conditions does not come close to human discriminability. Most importantly, however, the relative performance for the computational model closely follows human behaviour with increasing value of a , that is, inter>extra>ortho. For $10 < a < 100$, the relative change between the four conditions seems to correspond best to the human data, whereas for $a = 0$, this performance pattern is not yet as pronounced as for larger values of a . This result provides support for the existence of a horizontal matching prior, which produces this *qualitative shift in behaviour* as observed both in the computational and human data. Interestingly, the results for *inverted* faces (labelled HumanInv) in which inter>extra=ortho was observed, correspond closely to the computational data for $a = 0$ which exhibits the *same* overall behaviour. The proposed computational model would therefore be able to explain human data, if the horizontal prior would *not be active* for inverted faces (in other words, if $a = 0$).

Figure 4.2b) shows the results for running the experiment with 20 faces instead of 10 faces. A comparison with the results of Figure 4.2a) shows that performance in all conditions stays well within the variability of the 10 face-data. Doubling the training and testing complexity therefore does not have any adverse effect on performance, nor does it change the pattern of results.

Finally, Figure 4.2c) shows the results for the baseline condition in which recognition was done using simple Euclidean distance between two images rather than using local features. First of all, recognition performance is much worse than both human and local feature performance. In addition, there seems to be a large drop for the ortho(up)-portion of the viewing sphere, resulting in a large asymmetry for the ortho condition, which cannot be seen in either the upright or the inverted condition in the human data. Interestingly, there still seems to be a tendency towards inter>extra>ortho for this recognition algorithm. Repeating the experiment with 20 faces, however, results not only in a noticeable overall drop in performance but also in a *qualitative* change towards inter>extra=ortho(down)>ortho(up). Given this instability for an increased learning set and the asymmetry in the ortho condition as well as the inferior performance compared to the local feature model, it seems that using holistic Euclidean distances for recognition does not adequately describe human performance for neither upright nor inverted face recognition. In contrast, a lo-

Figure 4.2: Total recognition performance as AUC-values in the inter, extra, ortho(up) and ortho(down) conditions as a function of the horizontal viewing prior. a) Local feature matching with 10 faces, b) local feature matching with 20 faces. The first four bars in each graph correspond to a value of $a = 0, 10, 100, 1000$ respectively, the last two bars in the right graphs show human performance for both the upright and inverted condition. c) Baseline condition consisting of L2-based matching with 10 and 20 faces, respectively.

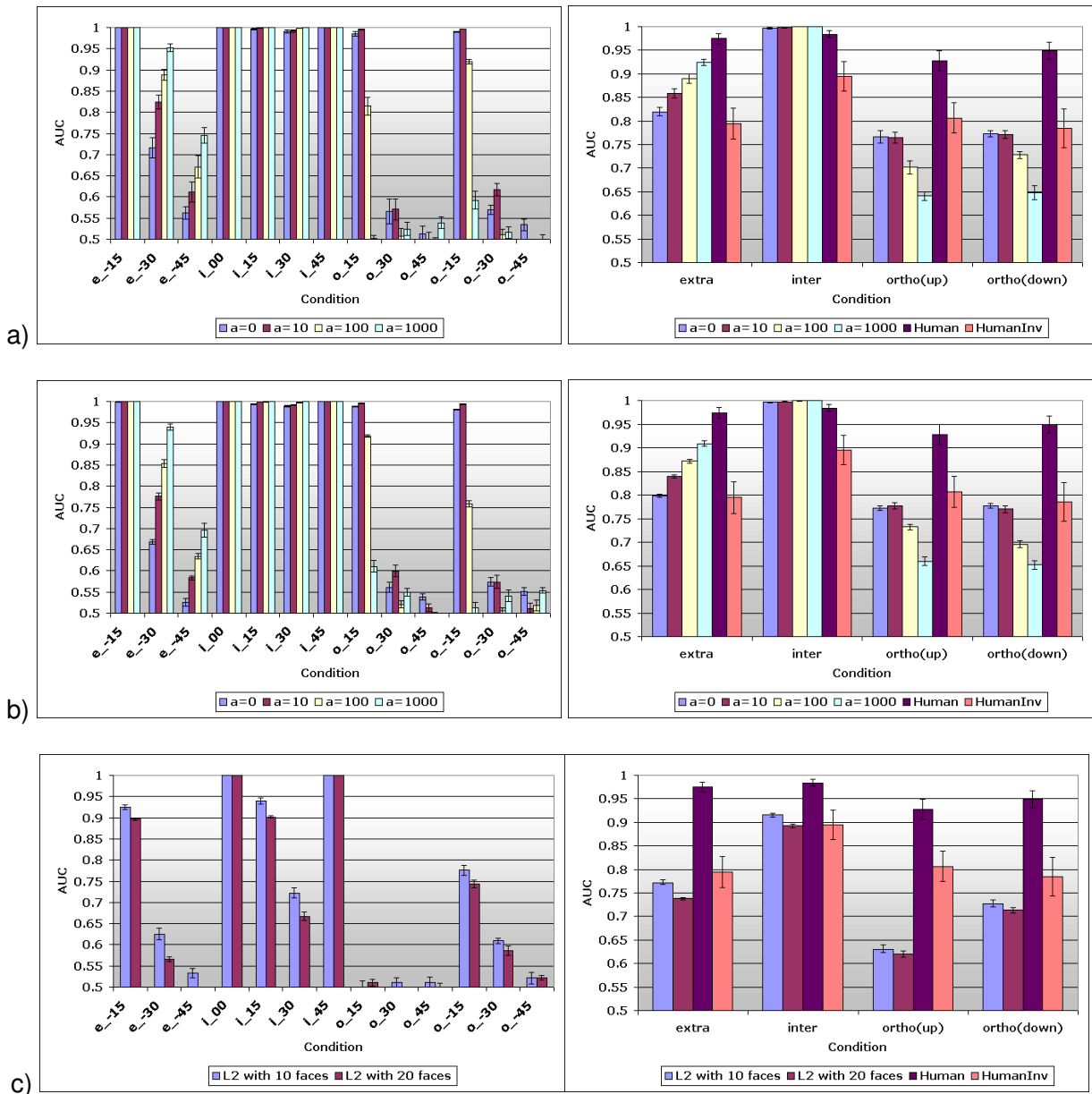
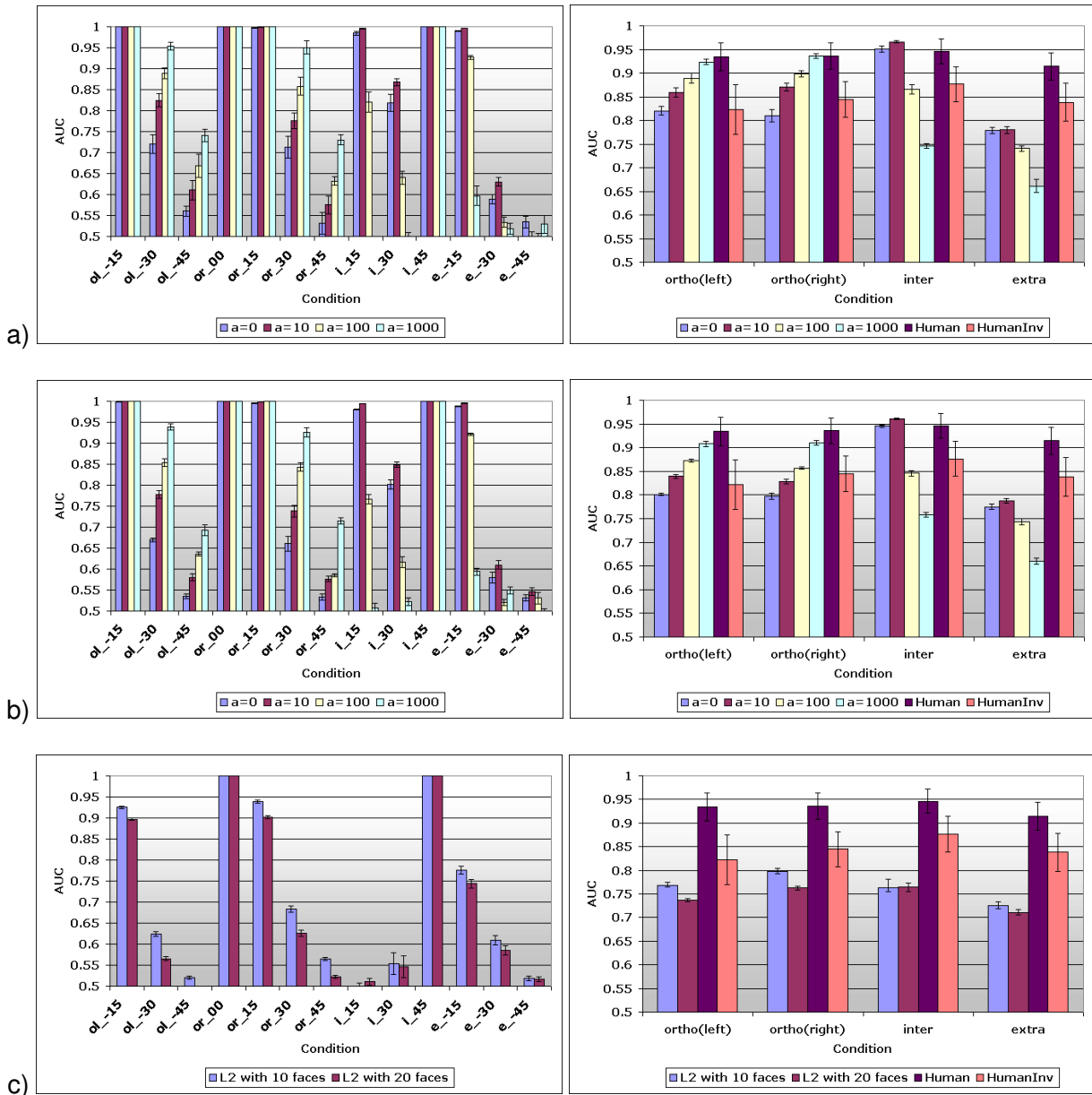


Figure 4.3: Total recognition performance as AUC-values of recognition performance in the inter, extra, ortho(left) and ortho(right) conditions as a function of the horizontal viewing prior. a) Local feature matching with 10 faces, b) local feature matching with 20 faces. c) Baseline condition consisting of L2-based matching with 10 and 20 faces, respectively.



cal feature matching scheme for increased performance and stability coupled with a horizontal matching prior describes the observed performance pattern surprisingly well.

Experiment Design: For the second experiment, the system was trained with two images of 10 target faces on the *vertical* viewing axis (0° and 45°). All other experimental settings were the same.

Results+Discussion: Figure 4.3 shows the results from computational experiments as well as a comparison to human data (note that the ortho conditions now refer to the *horizontal* viewing axis, whereas both the inter and extra conditions refer to the *vertical* viewing axis).

First of all, the horizontal viewing prior results in a different performance pattern with increased discriminability in the two ortho conditions at the expense of both the inter and the extra condition. For $a = 0$, the inter condition performs best, followed by the two ortho and finally the extra condition. For larger values of a , this pattern changes until for $a = 1000$, recognition is best in the two ortho conditions on the horizontal viewing axis, followed by the inter and the extra condition. As in the previous experiment, there is no significant difference for the increased learning and testing set with 20 faces compared to only 10 faces.

Comparing this to human data, where $\text{inter}=\text{ortho}>\text{extra}$ was shown, the values for the matching prior which produce the same qualitative behaviour are in the range of $10 < a < 100$ - the *same range* as found in the previous experiment. Similarly, $a = 0$ corresponds more closely to the human data in the *inverted* condition with $\text{inter}>\text{ortho}=\text{extra}$, which provides further evidence for the hypothesis that the matching prior is only effective for *upright* faces.

Finally, results for the Euclidean matching follow a similar pattern as before: first of all, performance is much worse than for the local feature matching. Second, performance changes noticeably between the 10-face and 20-face condition, showing an inferior generalization capability with respect to local feature matching. Finally, the observed performance pattern changes from $\text{ortho}(\text{right})>\text{ortho}(\text{left})=\text{inter}>\text{extra}$ to $\text{inter}=\text{ortho}(\text{right})>\text{ortho}(\text{left})>\text{extra}$ - the latter being closer to human performance in the inverted condition. Nevertheless, the observed instability together with the inferior performance compared to local feature matching suggest that, again, Euclidean matching provides a poor model of human performance.

4.1.3 Summary

As was shown in the two computational experiments, local feature matching in combination with a horizontal matching prior seems to model human performance surprisingly well. Both the horizontal as well as the vertical training results have suggested the *same* range of values for the matching prior. In addition, local feature matching provides a much better performance, generalizability as well as robustness compared to a simple, Euclidean matching strategy. This result in particular suggests that image-based, *holistic* algorithms need more sophistication until they achieve the performance level of image-based, local algorithms - let alone human performance. It would be interesting, for example, to couple the holistic image representation with view interpolation techniques or a sophisticated statistical learning algorithm in order to see whether this might improve the correlation with human data.

Furthermore, the computational experiments have shown that if the influence of the prior is reduced to $a = 0$, the observed performance pattern closely matches human data in the *inverted* condition - both qualitatively and quantitatively. This suggests that in humans, the matching prior might only be present for upright faces. A straightforward explanation for this difference in processing strategies could be the viewing statistics, which favor views of upright faces along the horizontal axis and which therefore reflect prior experience or expertise with this class of stimuli. This explanation, however, deviates from the *general* assumption of a horizontal viewing prior put forward by Bülthoff and Edelman [1992]. Such a discrepancy might simply result from the difference between processing of novel and familiar objects - in particular, upright faces are processed differently than inverted faces, both of which are processed differently than novel objects. Further

experiments - both psychophysical as well as computational - are needed to investigate these issues in more detail.

Finally, I want to briefly discuss the relation to the configural and component processing that was discussed in chapter 1. As was shown, the distinction between these two types of processing is able to explain such diverse results as the Thatcher illusion, the Hayes-Young illusion as well as a number of other psychophysical results Schwaninger et al. [2003]. In the computational experiments in this section, I have used structural elements of the configural and component route, that is, the existence of a geometric or spatial analysis as well as the processing across two different spatial scales (a fine scale for component information as well as a coarse scale for configural information Goffaux et al. [2005], see also next section for an in-depth discussion of configural and component processing). Interestingly, a closer look at the recognition data shows that the coarse scale information is much more important for recognition across pose changes than the fine scale information. This in turn would lead to the prediction, that, if one were to combine this IEO experiment with the scrambled-blurred experiment, one should see a much more *reduced* performance of component processing in the scrambled condition in contrast to a *reasonably good* performance of configural processing in the blurred condition. This experiment will be conducted in the near future, thus closing the loop between psychophysical experimentation and computational modeling.

4.2 Configuration and components

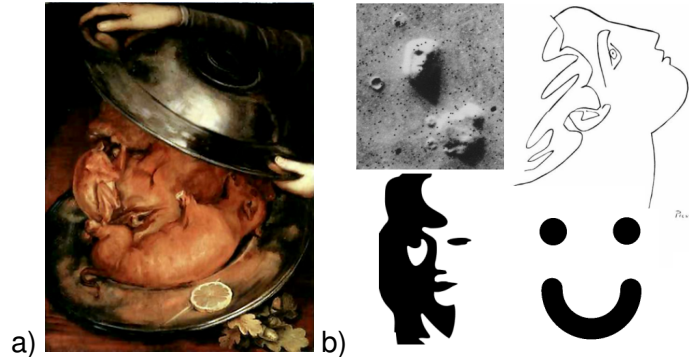
This section presents modeling results, which rely mainly on the *local feature matching* component of the keyframe architecture. As was discussed in chapter 1, psychophysical results indicate that face processing relies on processing of holistic and configural information - on the whole and its constituents. Given this evidence, I will first introduce a possible implementation of such a processing scheme and validate its performance on the same set of stimuli. In addition to the modeling results, first recognition results from computational experiments show that by using such a structured representation, recognition under larger view rotations becomes more robust.

As was shown in chapter 1, face processing seems to rely on relations between features - parts and their configuration. The question is now how to implement these two processing streams in order to be able to model the psychophysical results. In my view there are two important aspects in this respect: the first is given by the fact, that in order to enable processing of parts and configurations, some sort of *local* processing must take place. This connects the proposed psychophysical model (Figure 3.1) with the computational framework detailed in the previous chapter as well as the discussion and modeling of view-based recognition in the previous section. Of course, "parts" are more than just a few local features - they are highly salient for robust recognition of exemplars; in addition, they usually are class-specific in that they can for example be determined by maximizing within-class similarity while minimizing between-class similarity (see for example, Edelman [1999], Graf et al. [2002], Ullman et al. [2002]). The second important aspect complements the previous one: parts rely on *relations*, which result from some form of geometric analysis. It is not only the parts themselves, which allow categorization for example, but also their configuration, their geometric layout. Faces are perhaps the most convincing argument for this, as the example in Figure 4.4a) shows and as is also demonstrated by the ease with which extremely sparse information is interpreted as containing faces (Figure 4.4b). In the following, I will focus mainly on faces based on the psychophysical experiments in chapter 1, although the algorithms could in principle be also be applied to other object classes.

4.2.1 The face representation

To support the two aspects of salient local features in a relational or geometric context, I propose to process a face image in a manner similar to the one described in the previous chapter: First

Figure 4.4: Illustration of the importance of geometric information (configuration) in interpretation of images. a) Painting by Arcimboldo ("The Cook", 1570) - turn upside down for a second interpretation! b) Collection of images showing how little visual information is necessary to interpret something as a face.



of all, a visual representation is extracted based on a set of interest points, which are determined at several scales. To each interest point location a small neighborhood of 5x5 pixels as appearance information is added to form a local visual feature - this constitutes one of the most basic appearance-based local object representations in the context of the definitions given in chapter 2. Figure 4.5 shows reconstructions of two faces from such a feature representation, in which features from coarse scales were resized according to the scale difference and then images from all scales superimposed starting with the coarsest scale (see also Figure 3.2). Two aspects are worth noting here: First, even though the representation is quite sparse in terms of compressing the original data (a total of 160 features each of which contains 25 pixels results in a compression rate of 93.9%), the reconstruction still gives a fairly good visual impression of the original face. Second, and perhaps more importantly, one can see that the extracted features tend to cluster around important facial features (see especially Figure 4.5b). Eyes, mouth and nose are represented with a much higher density of features than, for example, the forehead or the cheeks. Note that the claim here is *not* that a simple corner detector is able to explain the complicated processes of feature formation, as there are of course other types of information available for this (one obvious cue might, for example, be facial motion resulting from facial expressions or talking). It seems, however, that these basic appearance-based visual features already capture some of the essence of "semantic" facial features in the chosen setting.

In addition to these image fragments, *geometric information* is captured by extracting for each visual feature a vector containing relative pixel distances to a number of neighboring features. This "distance vector" is used during the matching stage to determine either the component or the configural properties of each feature. In order to facilitate processing, the distance histograms are sorted in *increasing* order (smaller distances to larger distances). For each point i , the distance vector \vec{d}_i thus is determined as:

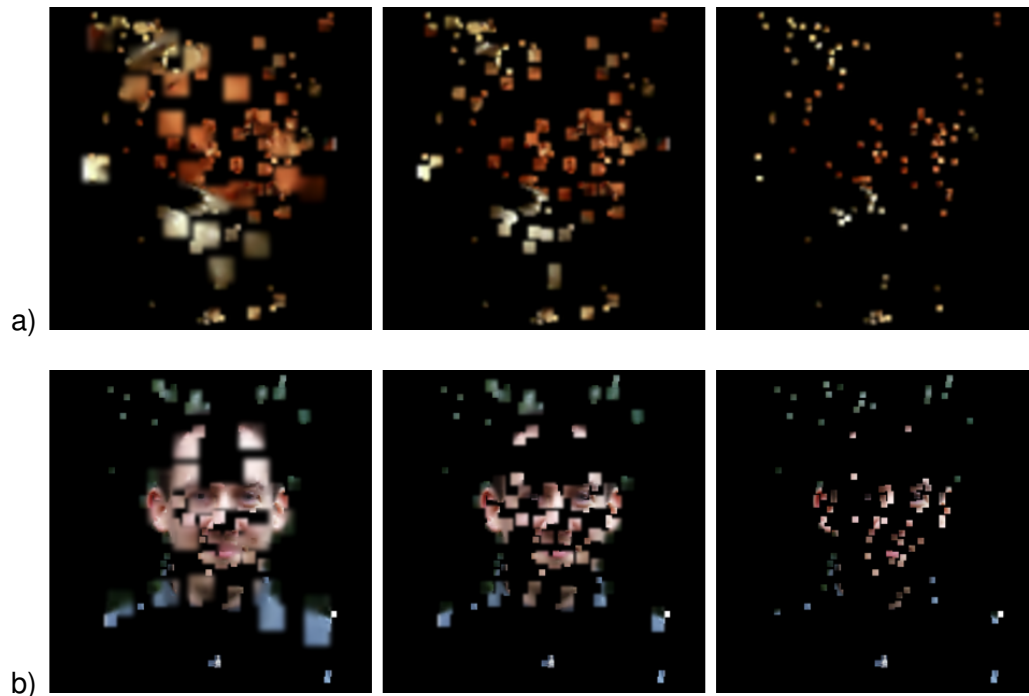
$$\vec{d}_i = \text{sort} \left\{ \frac{1}{\text{size}(I)} \|\vec{p}_i - \vec{p}_j\|^2 \right\}_{j \neq i}$$

The factor $\frac{1}{\text{size}(I)}$ is used to normalize all distances with respect to the current image dimensions (which are dependent on the level of the Gaussian pyramid, see also chapter 3).

From this definition it is easy to see, for example, that \vec{d}

- characterizes the local as well as the global aspects of feature layout
- takes on the same values for a uniform (in terms of distance in the image plane) feature

Figure 4.5: Reconstruction of images of faces from the visual feature representation. Note that features tend to cluster around important facial features such as eyes, mouth, nose, etc.



distribution

- is invariant to image rotations and scaling that preserves the aspect-ratio of the image (provided that feature extraction is invariant to these image manipulations)
- is *not* invariant to more general affine transformations, which, for example, might affect the x and y-axis of the image differently

Using these two basic building blocks of local visual features and their image relations, I now want to outline a method of defining component and configural processes. First of all, the proposed method does not use prior knowledge about facial parts, which would for example be available in the form of state-of-the-art facial feature detectors (see Hjelm and Low [2001] for a recent overview) or the use of a sophisticated 3D face database (such as Blanz and Vetter [1999]). Instead a purely bottom-up data-driven definition of such "parts" was chosen that can accommodate different object classes but at the same time is flexible enough to allow later learning of a more abstract definition of parts. Parts in this manner can be defined using the two types of extracted information as *tightly packed conglomerates of visual features at detailed scales*. As will be seen in the computational experiments, this definition captures the most important aspects of the model defined in section 1.2.6 *without* using prior knowledge in the form of pre-learned part models (such as templates for the eyes, or spline models for the mouth). It should be noted, that by construction, parts in the framework are defined by a form of *configuration* in the feature set. This implies that their processing relies on the relationship between features at detailed scales. Complementing this processing strategy configural processing can then be defined based on the *relationship between visual features at coarse scales*. These two processing strategies, however, will be made explicit only during the feature matching process outlined below. Interestingly, a recent psychophysical study has provided strong support for this separation into coarse and fine scales by showing how configural processing relies on low spatial frequencies (corresponding to the coarse levels of the Gaussian scale pyramid), whereas component information relies on

high spatial frequencies (corresponding to the fine levels of the Gaussian pyramid) in the image [Goffaux et al., 2005].

One view of component and configural processing would be that configural processing relies on the relationship between the extracted parts. This would have an important consequence for the processing of faces, namely, that part detection comes first followed in a second step by configural processing. Equally valid, however, would be the assumption that the configural route is activated first by a coarse configural description of the stimulus, which then is able to trigger a more detailed processing of parts. The advantage of the second strategy lies in its coarse-to-fine processing, where information from configural processing can be used for later matching of parts. As the psychophysical experiments show, however, humans seem to be able to do part-matching without the help of configural processing with a reasonable performance (see section 1.2.6). This means that although these two routes might complement each other under "normal" circumstances (that is, for recognition of intact faces), the experimental evidence also supports two more independent processing structures. In the following, I will thus adopt the coarse-to-fine strategy in the form of the multi-scale image representation, but I will keep the two types of processing largely separated.

4.2.2 Feature matching

The algorithm for recognition of (face) images is the second main part of the computational modeling of configural and component processing. As each image consists of a set of visual features, recognition amounts to finding the best matching feature set between a test image and all training images.

The two routes for face processing are reflected by two types of matching algorithms based on configural and component information and is an extension of the matching equation presented in section 3.4.2. First, a similarity matrix \mathbf{A} is constructed between the two sets, where each entry $A(i, j)$ in the matrix is:

$$A(i, j) = \exp\left(-\frac{1}{2\sigma_{\text{geo}}^2} \text{confcomp}(\vec{p}_i, \vec{p}_j)\right) \cdot \exp\left(-\frac{1}{2\sigma_{\text{app}}^2} \text{NCC}(\vec{l}_i, \vec{l}_j)\right)$$

The second term is the appearance similarity measure given by the normalized gray value cross-correlation between the two pixel patches i and j . The first term, however, now reflects the two types of processing routes. Given the distance vectors for features i and j , the function confcomp is evaluated as:

$$\text{confcomp}(\vec{p}_i, \vec{p}_j) = \sum_{k=1}^N \|\vec{d}_i(k) - \vec{d}_j(k)\|$$

Component matching is done in this framework by restricting N to the first few elements of the distance vector \vec{d} , thus restricting analysis to close conglomerates of features - a *local* analysis. Configural matching on the other hand relies on *global* relationships, such that N is restricted to the last elements of the distance vector. The size of N should be small for component matching (in the following experiments, $N=3$ was used) and larger for the global configural matching (in the following experiments, $N=|d|$ was used). Note that for a given feature location, a large value of N biases confcomp towards the extreme, far away features. In addition, the parameters σ_{geo}^2 and σ_{app}^2 can be used to control the relative importance of the two types of information. Note also, that the difference between this and the approach used in the previous section lies mainly in the use of another geometric similarity term.

The matrix \mathbf{A} thus captures similarity between two feature sets based on a combination of image-based feature layout information and appearance information. As outlined in section 3.4.2, corresponding features can now be found with a greedy strategy by looking at the largest elements of \mathbf{A} both in row and column satisfying $A(i, j) > \text{thresh}_{\text{app}}$, which yields a one-to-one mapping of one feature set onto the other. The threshold is used to introduce a quality metric for the matches

and is in this case applied to the appearance similarity measure only. The percentage of matches between the two feature sets for the component route and the configural route then constitute the final matching scores. *Adding* these two scores in order to integrate the two routes then yields a total matching score - this corresponds to a simple, linear cue integration model.

4.2.3 Modeling psychophysical experiments

In this section, I will describe the computational modeling experiments, where the implementation was applied to the psychophysical experiments from chapter 1 using the *exact same stimuli*. For a second set of experiments in which it was investigated to which degree the two separate routes for recognition would be beneficial also for other recognition tasks, see section 5.1.

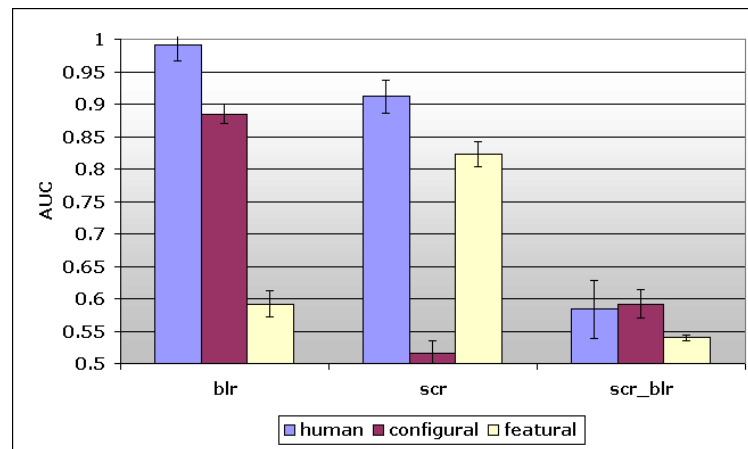
Experiment Design: For the computational modeling experiments, a total of 20 faces from the psychophysical experiment were used. Apart from the original image, each face was available in three versions: scrambled (Scr), blurred (Blr) and scrambled/blurred (ScrBlr). The experimental protocol was as follows: For each run of the experiment, 10 faces were selected as target faces and 10 faces as distractors. The 10 target faces were learned by first extracting visual features and the distance vectors as outlined in the previous section and subsequent storing of these representations. In the testing phase, a face image was presented to the system, which again extracted the feature representation and then found the highest matching score among the 10 learned faces using the two-route matching procedure. In order to better judge the influence of the two routes, each route was run separately on the data *without* adding the final matching scores. The presented face could be either a target or a distractor in one of the three stimulus versions. In order to get a better statistical sampling, the experiment was repeated 10 times, each time with a different set of target and distractor faces. In a next step, the experimental data was converted into ROC-curves by thresholding the matching scores for the target faces and the distractor faces, respectively. The area under the ROC-curve was taken as a non-parametric measure of recognition performance (again, $0.5 \leq \text{AUC} \leq 1.0$) that can be compared directly with the psychophysical data.

From the description of the implementation in the previous section it is clear that a number of internal parameters can affect the performance of the system in the various experimental conditions. The first parameter is the number of features, which specifies the complexity of the data representation. One might expect, for example, that adding more features to the representation would lead to better overall performance. The second set of parameters is given by σ_{geo}^2 and σ_{app}^2 , which control the relative importance of appearance and geometric information. The third, related, parameter is the quality threshold $\text{thresh}_{\text{app}}$, for which one might expect that with increasing threshold the discriminability of the found matches will also increase. These parameters allow us to characterize the parameters of the system with respect to the human performance data obtained in the psychophysical experiments.

Results and Discussion: Figure 4.6 compares AUC-values for human data with AUC-values for the computational implementation. In addition, the computational data are separated to show the contributions of the configural route and the component route in the different conditions. First, it can be seen that the computational performance is slightly lower than the human performance. This can be attributed to the simple visual features that were used in our implementation - more sophisticated visual features such as the ones developed in Lowe [2004] could provide better recognition performance.

More importantly, however, the relative contribution of the two processing routes follows exactly the expected pattern with the configural route being active in the blurred condition and the component route being active in the scrambled condition. In addition, the configural route does not contribute to recognition in the scrambled condition; similarly, the performance of the component route in the blurred condition is negligible. Performance of both routes reaches chance level in the scrambled and blurred condition. In addition, the relative contributions of each route closely follow

Figure 4.6: Unfamiliar face recognition: AUC values for all conditions of the human data, the computational data split into contributions by the component (featural) and configural processing route, as well as standard local feature matching (std). The error bars depict SEM.



the human data.

Figure 4.6 also shows the results of running standard local feature matching *without* the geometric constraint on the stimuli. Whereas there is no difference for the scrambled stimuli (which is not surprising, given that both algorithms are virtually identical), recognition performance in the blurred condition drops to the level of performance in the scrambled condition. This result demonstrates that the additional geometric constraint not only helps to increase recognition performance but that it seems *necessary* for this local feature matching framework in order to be able to capture the relative performance difference of the human data.

Taken together, this pattern of results models the pattern observed in the psychophysical experiments on a qualitative level and thus provides initial evidence for the perceptual plausibility of our implementation of the two routes of visual processing.

In Figure 4.7, an example of feature matching in each of the three conditions is given - corresponding features are indicated as white dots. In this example, the component route is active for the scrambled condition, the configural route for the blurred condition, whereas only one match could be found in the scrambled and blurred condition. The full experimental results in Figure 4.6 confirm that both routes process the information independently as AUC-values are negligibly small for the conditions in which only one type of information should be present. In addition to the quantitative results and the relative activation of the two routes in the different condition, this provides further evidence for the plausibility of the implementation.

Furthermore, Figure 4.7 shows that part-based matching concentrates on high-level details such as corners of the mouth, points on the nose, some features in the eyes and on the eyebrow, etc. Interestingly, this observation already leads to concrete experimental predictions, which can be used for further psychophysical studies: most of the matching features in the component routes focus on high-contrast regions (due to the nature of our visual features). If component processing in humans relies on similar low-level information parts in this experiment, which have less high-contrast regions (such as the forehead or the cheeks) should contribute *less* to the human recognition score. It would be interesting to design a psychophysical experiment, which directly addresses this question of how different parts are weighted in recognition - this would represent a good example of how computational modeling could feed back into cognitive research. Configural matches, in contrast to component matches, are in general spread much further apart in the image (tip of the nose, nose bridge, features on the cheek), which carry less appearance information but are globally consistent local features in terms of their spatial layout in the face. Nevertheless, it has

Figure 4.7: Corresponding features for the three test conditions (upper row: blurred, middle row: scrambled, lower row: scrambled and blurred). Component route matches are shown in white, configural route matches in cyan - the lines connect corresponding features *in the image plane*.

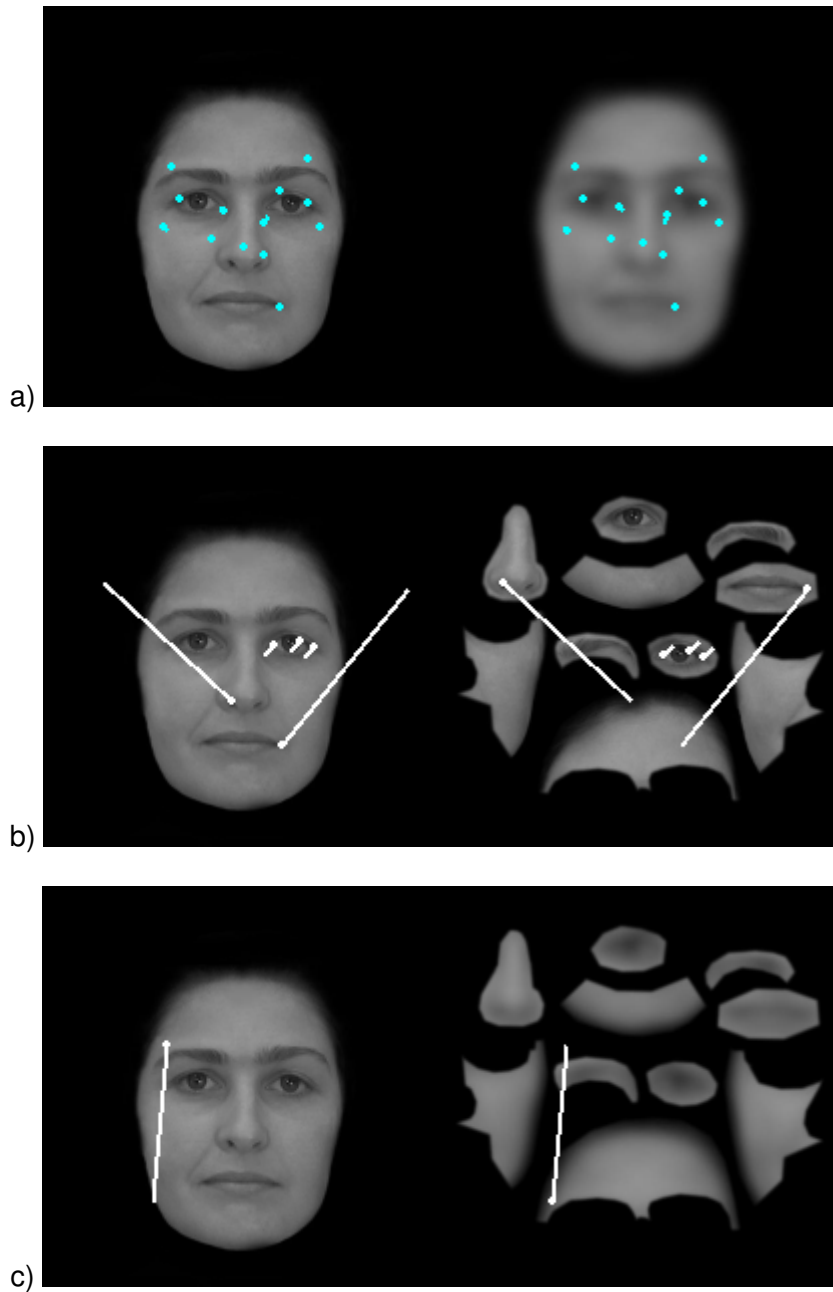
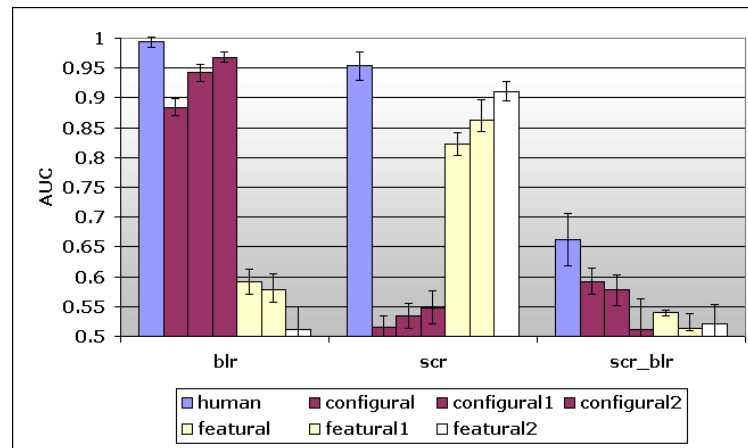


Figure 4.8: Familiar face recognition: AUC values for all conditions of the human data as well as the computational data split into contributions by the component (featural) and configural processing route. Computational data is based on three visual representations with increasing visual complexity. The error bars depict SEM.



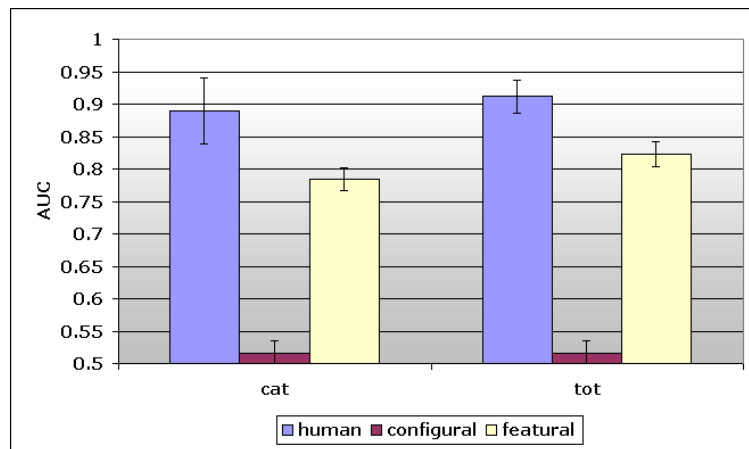
to be stressed that the locations of the configural features are also determined by high-contrast regions - albeit on a lower spatial scale. As shown in Figure 4.7a+b, there is also some overlap of feature locations across spatial scales which is in accordance with results from Maurer et al. [2002] who showed that configural and component processing share information.

Experiment Design: In a second step it was tested how well the computational implementation would be able to capture the effects of familiarity observed in the psychophysical experiments. One of the most obvious parameters that might be responsible for the difference between familiar and unfamiliar face recognition might be the richness or complexity of the extracted representation. If humans are repeatedly exposed to the same face, this experience could simply result in a more detailed representation of its visual appearance. The computational counterpart to this in our computational implementation would consist of the number of local features that constitute the representation of a face image. The following computational experiment explicitly tested this hypothesis with the stimulus set of the previous experiment by systematically increasing the number of features in each processing route.

Results and Discussion: Figure 4.8 shows AUC-values for the human data from Experiment 2 compared with AUC-values for the computational implementation. The computational data is shown for three different sizes of the visual representation: original (same as in the previous experiment), the number of local features increased by 50 percent, and the number of features increased by 100 percent. As hypothesized, the performance of the computational data increases with increasing visual complexity in both routes. In contrast, the results for the configural route in the scrambled condition and for the component route in the blurred condition show no systematic increase with increasing visual complexity. Most importantly, the relative contribution of each route does not change in the three conditions. In addition, the performance of the most complex visual representation approaches human performance - a further increase in number of features, however, does not provide better recognition performance, indicating that the discriminatory power of the simple visual features used in this study has reached its limits. The experimental results presented here suggest that a surprisingly simple parameter such as the complexity of the visual representation might be sufficient to explain the increase in performance observed in the psychophysical experiments.

Experiment Design: Whereas in the previous two computational experiments we were interested to model unfamiliar and familiar face recognition, in this experiment we aimed at reproducing the

Figure 4.9: AUC values of scrambling condition for human and computational data where categorical spatial relations are left intact (Cat) versus where they are totally scrambled (Tot) - see also Figure 1.15. The error bars depict SEM.



independence of scrambling type found in Experiment 3 in the original psychophysical study (see 1.15b). The computational experiment was therefore repeated with the same set of categorically scrambled stimuli and compared with the results from the non-categorically scrambled face images used before.

Results and Discussion: The results of this computational experiment are shown in Figure 4.9 for the two types of scrambling (Cat and Tot). Similarly to the human data, the computational performance remains unaffected by type of scrambling used, thus providing further support for the plausibility of our implementation. This is confirmed by a two-sample t-test (two-tailed), which yields *no significant difference* between the two conditions for the component processing route, $M=0.82$, $t(11)=1.46$, $p=0.16$.

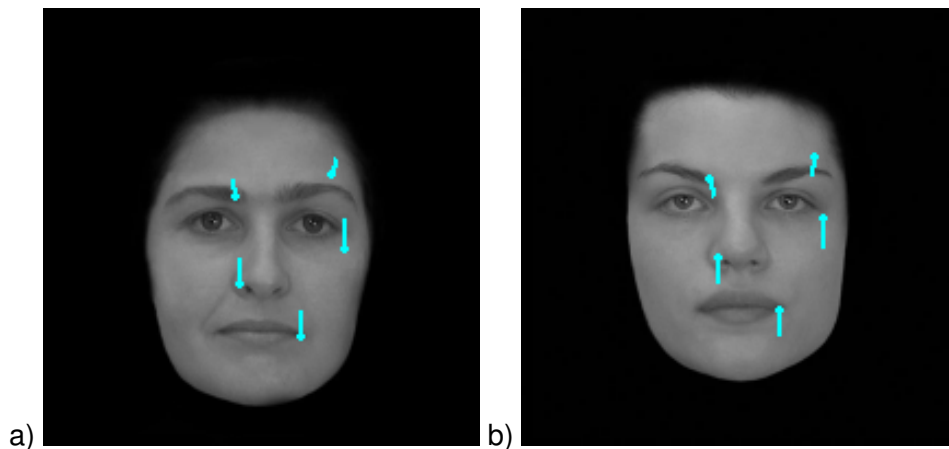
General Discussion: In summary, our results show that our implementation of the two-route architecture is able to capture the range of human performance observed in the psychophysical experiments. In addition, changes in the internal parameters of the architecture - we have so far investigated visual complexity and discriminability - result in plausible changes in observed performance while retaining the overall qualitative similarity to the human data in terms of the observed weighting of the two routes. Finally, several observations from the computational experiments can be used to plan further behavioral experiments, which will investigate the features of face recognition in the context of the two-route architecture.

The fact that configural processing is largely based on spatial properties face also allows for *categorization* of faces, as the configural information captures the global layout of face structure. Figure 4.10 shows an example of the full matching result between two faces, which demonstrates the generalization capabilities of our system². Again, the only active route in this picture is the configural route - no matches from the highly detailed component route were found in this image.

In the following, I want to discuss the implementational details of the proposed architecture in more detail from a computer vision perspective. From this point of view, one might ask why the component route, for example, does not use established computer vision procedures and representations such as template matching, cross correlation techniques or even histogram matching instead of the visual features introduced in section 3.4.1. With all of these standard approaches, one should be able to recover nearly all component matches perfectly, as the scrambled condition consists basically only of a permutation of the original pixels. From a cognitive perspective, however, there is evidence that humans do not seem to pay attention to images at the level of *single*

²In order to achieve this results, the matching threshold was decreased compared to the previous experiments.

Figure 4.10: Corresponding features for two faces showing that the general class-based similarity in layout is captured well by the configural route in our implementation. a) original face with feature displacements b) test face with matching features.



pixel information [Rensink et al., 1997] (as would be the case for the computer vision techniques). Humans rather seem to rely on a more abstract representation of visual data - maybe even including a semantic representation such as "full mouth", "curved eyebrows", which is based on a higher-level interpretation of the visual information. Although the proposed representation is not semantically grounded, it is extensible to a semantic and thus class-specific representation (see, for example, Ullman et al. [2002], Leibe and Schiele [2003] for approaches in this direction).

Alternatively, one might also base the need for a more abstract representation on memory or storage constraints: the amount of visual memory necessary to save this kind of detailed pixel information is simply not available for this task. The proposed implementation of the two processing routes can be seen as an embodiment of such a memory constraint: the huge number of possible visual features and their image relations is reduced to a few of the most salient ones taking into account their local neighborhood for a larger number of detailed features and their global neighborhood for a smaller number of coarse features. Whereas from a computer vision perspective the task itself could be solved with almost perfect recognition performance - even though at a significantly higher memory load - the extraction of visual features enables a much sparser and more abstract representation. In addition, their inherent robustness allows for extraction of further abstract information - such as analysis of visual features across all learned faces to extract parts and common feature relations, etc. In summary, apart from providing one layer of data abstraction, our implementation of the visual features underlying the two processing routes thus seems to be able to fit well into models of human visual memory.

4.2.4 Summary

Psychophysical evidence strongly supports the notion that face processing relies on two different routes for configural information and component information. We have implemented a simple computational model of such a processing architecture based on low-level features and their 2D geometric relations, which was able to model the psychophysical results in a qualitative manner. In this context it has to be said that an exact quantitative modeling - while this might seem a desirable goal - cannot be realistically achieved as there are too many hidden variables in the exact formation of the psychophysical data. Some examples include the different contexts of human and computational studies: while humans have a life-long experience with faces and can use that to encode the faces in both training and test stage, the computational system does not

use any prior knowledge about faces. The plausible behavior that was observed in changing the internal parameters of the implementation, however, seems to be a good indication that human performance in the psychophysical task and our implementation of the computational architecture share a similar structure on the *functional* level [Marr, 1982]. Further work in this area should focus on more class-specific processing such as learning of semantic, appearance-based parts from local features of faces (see, for example, Ullman et al. [2002], Weyrauch et al. [2004]) together with their spatial layout.

4.3 Temporal aspects of recognition

In this section, I will focus on the *temporal* dimension in object recognition, in particular on the morphing experiments outlined in chapter 1, which represented one inspiration for the keyframe approach in that learning and recognition of objects crucially depend on spatio-temporal aspects of the input data. In particular, Wallis and Bülthoff [2001], Wallis [2002] used sequences of rotating faces which either morphed during rotation or were even assembled from different faces as training sequences. In a later recognition task, participants had to decide whether two sequentially presented faces were from the same person. They found that participants confused faces more often in this task if they were contained in the training sequences than if they were not. In combination with the findings from several control experiments, these results indicate that learning of object representations is mediated by spatio-temporal continuity of images. The question I will address in this section is how to understand the effects of spatio-temporal continuity within the proposed framework and how it can be used to model some aspects of the psychophysical results found by Wallis and Bülthoff [2001], Wallis [2002].

4.3.1 Modeling temporal contiguity by learning keyframes

As was already mentioned in chapter 3, the keyframe representation relies on the idea of feature tracking over time in order to define a view-based image graph of the input sequence. If cast in this approach, learning an object representation of a rotating face as done in the psychophysical experiment by Wallis and Bülthoff [2001], Wallis [2002] would involve

- tracking of facial features across time for the sequence stimuli and
- the gradual build-up of a keyframe representation, which is terminated by the end of the sequence - several keyframes will be generated because tracking is not possible (neither does it seem plausible) over the full 180°.

Temporal continuity, of course, is then already included in the presupposition of a tracking process, which requires that spatial changes Δx are relatively small within a given time change Δt thus effectively limiting the image-plane speed of features $v = \frac{\Delta x}{\Delta t}$. There is, however, a second important aspect of the tracking process: in order to define *what features* are tracked over time one needs to be able to determine feature similarity³. This aspect becomes central when investigating the psychophysical experiment, as for the morphing stimuli used in the study, not only the position of a feature changes over time, but also its visual appearance due to the morphing process. Given the psychophysical evidence one can assume that a dramatic change in visual appearance causes the temporal association process to break. This aspect can be modeled by introducing an additional feature property $\frac{\Delta a}{\Delta t}$, which specifies the change in appearance from one time instant to the next.

³Note that this view of tracking is slightly different from approaches using banks of filters tuned to different speeds, where the image data is analyzed with respect to a certain range of speeds rather than with respect to a certain range of feature similarities as is done in this thesis.

It thus seems reasonable to assume that in order to form a consistent object representation over time, the learning process is characterized by two *parameters* specifying the constraints that can be tolerated by the visual system before the spatio-temporal association process fails. In the context of the matching equation presented in section 3.4.2, these are:

- a parameter for the speed (σ_v) of tracked features in the image plane and
- the maximum change in visual appearance (σ_{app}) by which two features can differ in two consecutive images

The temporal association results observed in the psychophysical study can then be modeled in terms of the second main concept of the proposed recognition framework, namely the *linked* representation of keyframes. As shown in Figure 3.1, keyframes are created once the number of consistently trackable features drops below a threshold $thresh_k$. Thus, it is intuitively clear that - if compared to a normal sequence of the same face rotating - for a given set of parameters $\sigma_{\{v,app\}}, thresh_k$ a morphed sequence will result in less features which can be tracked over time. This is due to the fact that a higher strain is placed on the tracking process because of changes both in the shape as well as in the pose. On average one can therefore expect *more* keyframes to be created for morphed sequences compared to non-morphed sequences.

In addition, there are two interesting results from the follow-up experiment [Wallis, 2002], which used rotation sequences made up of different faces instead of morphs that I want to investigate more closely. The first result was that the temporal association effects were *reduced* when different faces were used to substitute the morphs. Second, there were virtually no effects found for a scrambled presentation of images, which destroyed a consistent rotation interpretation but otherwise should have left the pure temporal *contiguity* intact. Both results can again be interpreted in a tracking framework as effects of the appearance and velocity parameters, as

- morphing may cause a “smoother” visual transition between two consecutive images than scrambling thus making temporal association easier
- corresponding features between two images in both scrambled conditions may be too far apart to fully support an association process

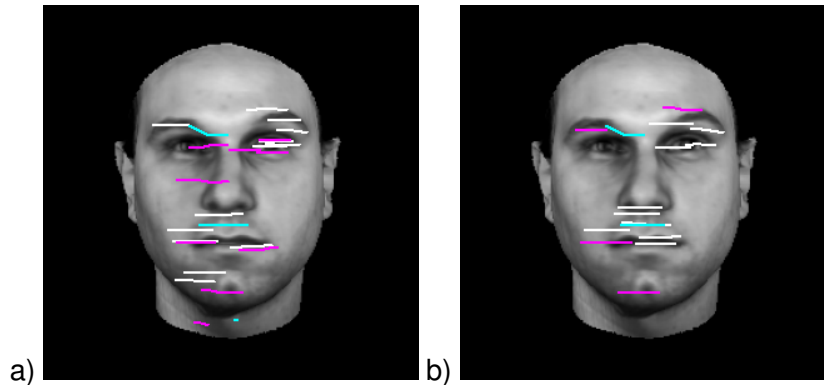
In both cases, as the number of frames is constant (and assuming that the thresholds stay constant as well), the number of keyframes is *increased* and thus in turn the measure of temporal contiguity is *reduced*.

Both effects have been found in the psychophysical studies which indicates that they can be modeled within the proposed computational recognition system with the help of the two thresholds σ_v and σ_{app} . In the following section, I will investigate the influence of these parameters on learning of sequences similar to the ones used in Wallis and Bülthoff [2001], Wallis [2002].

4.3.2 The influence of morphing on feature tracking

Experiment Design: The first experiment I want to present uses sequences of faces executing a relatively small depth rotation of 30° , which was chosen as to still allow for easy local feature tracking. A total of 100 sequences of 25 faces executing 30 degree rotations from -90° to -60° , -60° to -30° , -30° to 0° , -15° to 15° , 0° to 30° , 30° to 60° , and 60° to 90° (each sequence consisted of 7.5° steps) was analyzed by tracking. In addition, each sequence displayed either a normal face or a slowly morphing face (the amount of morphing was set to 10 percent per 7.5° degree step). The sequences were processed as described in sections 3.3, 3.4 by extracting local features in the first frame and matching of these features across subsequent frames. All computational results in this section used the standard local feature matching outlined in section 3.4.2 for tracking. The *amount of continuous trajectories* from the first to the last frame of each sequence was determined and

Figure 4.11: Comparison of feature tracking for a) a standard sequence and b) a morphed sequence of a face rotating 30° .



represented the critical measure for this experiment. This was done as a first test to see whether there would be a difference in the amount of tracked features in the normal versus the morphed condition. This difference could then give rise to the performance pattern that was observed in the psychophysical experiments.

In addition to the morphing manipulation, the visual complexity of the representation was varied by changing the number of extracted local features in 14 steps from 135 features down to 35 features. Following the discussion in the previous sections, this was done in order to explicitly test whether the behaviour of the tracking would change differently for normal and morphed sequences as the visual complexity of the representation would increase.

Results + Discussion: As an example, Figure 4.11 shows a side-by-side comparison of feature trajectories found by the matching process for a $0^\circ - 30^\circ$ sequence. In the normal condition (Figure 4.11a), tracking resulted in 25 feature trajectories over all resolution levels whereas the morphed condition (Figure 4.11b) resulted in only 18 feature trajectories, showing that morphing, indeed, produces a marked reduction in tracking. From a computational point view this does not come as a surprise since feature changes in the morphed condition are caused both by depth rotation *and* the morphing process.

This result is confirmed in Figure 4.12a in which the average number of trajectories from all sequences shows a large decrease from normal to morphed condition (a t-test showed this difference to be highly significant with $p < 0.001$): with respect to the total number of features in the first image, normal sequences on average result in 10% tracked trajectories, morphed sequences in 7%. In addition, it is interesting to see that a linear decrease of the visual complexity of the face representations results in a linear decrease in tracked trajectories. This demonstrates that the tracking process degrades 'gracefully' as the representation consists of fewer and fewer features. Furthermore, the slopes of the two lines are slightly different with the normal condition resulting in a steeper decrease in trajectories than the morphed condition. Reading the graph from right to left, this means that as the visual complexity of the representation increases, tracking is able to pick up a slightly larger number of trajectories in the normal condition.

Figures 4.12b+c show the average number of trajectories broken down by viewpoint for both normal and morphed conditions. It can be seen that the sequences around the more frontal viewpoints have consistently less feature trajectories than the sequences near the profile view ($p < 0.01$ for morphed sequences, $p < 0.01$ for normal sequences). This effect could be due to the face-specific symmetry that causes a higher feature change per degree around frontal than around side views (one obvious example for this effect is that the ears appear and disappear around the central viewpoint). In addition, the difference between the viewpoints becomes slightly less pronounced as fewer features are used for tracking. Finally, there seems to be a slight, yet

Figure 4.12: Average number of trajectories for 30 degree tracking as a function of number of features (visual complexity of the face representation) a) comparing normal and morphed sequences, b) showing the results for normal sequences broken down by viewpoint, c) showing the results for morphed sequences broken down by viewpoint.

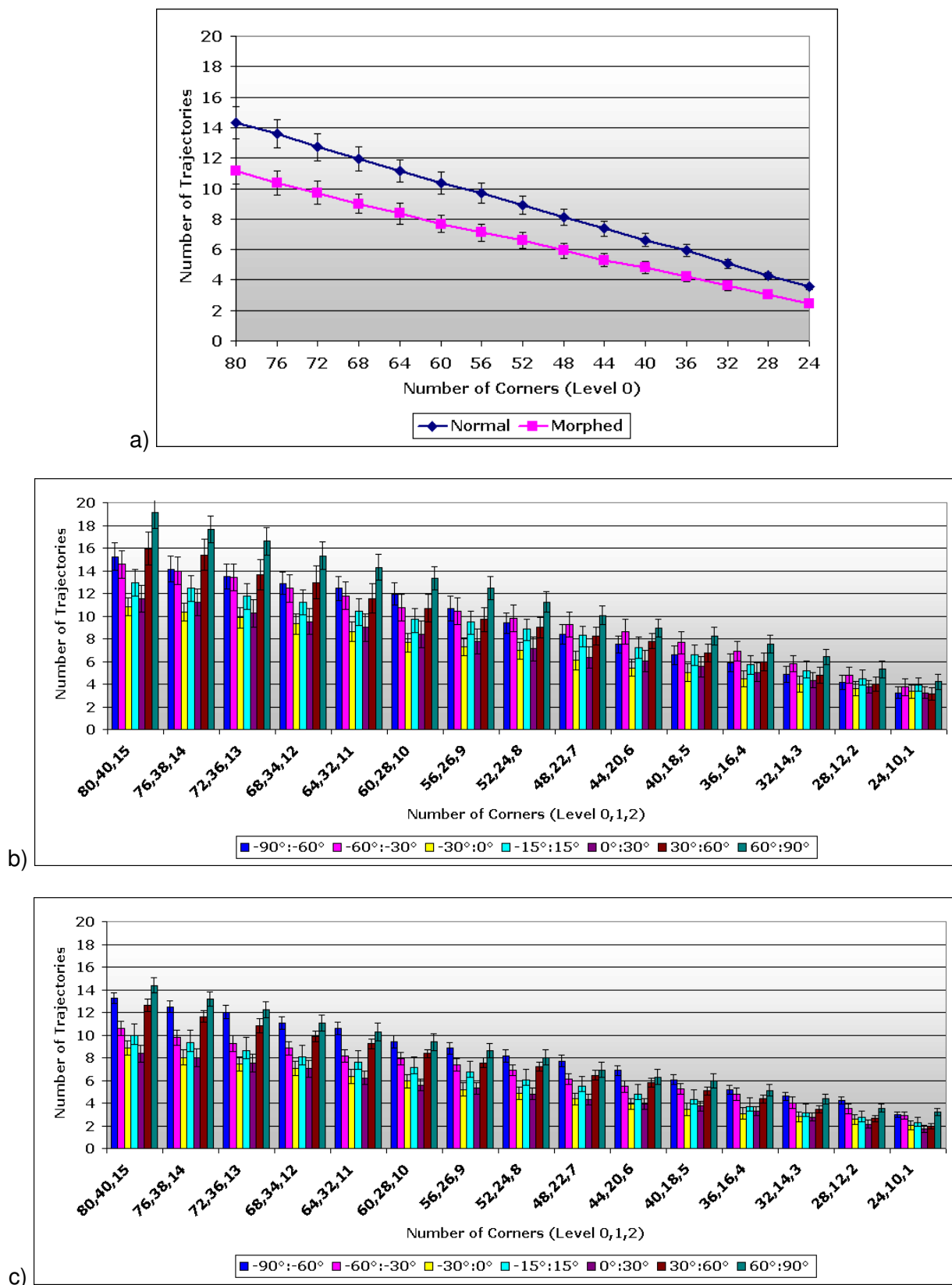
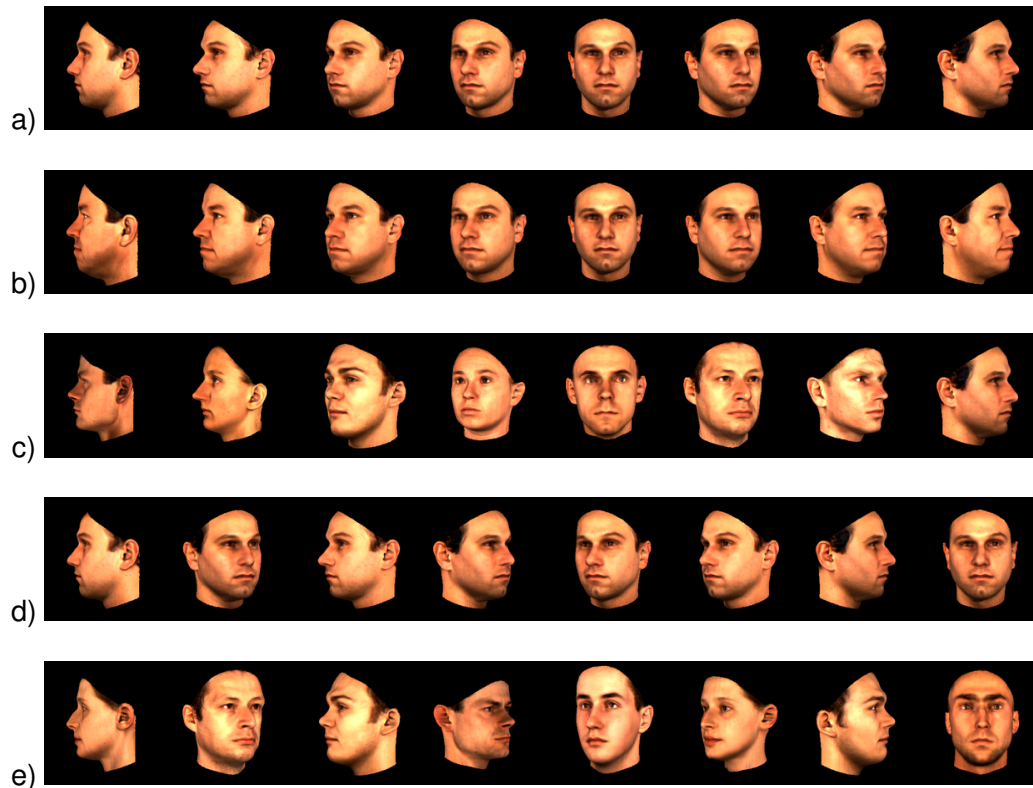


Figure 4.13: Example sequences for a) normal, b) morphed, c) identity scrambled (IdScr), d) pose scrambled (PoseScr) and e) fully scrambled (IdPoseScr) conditions. The figure only shows every third frame of the original sequence.



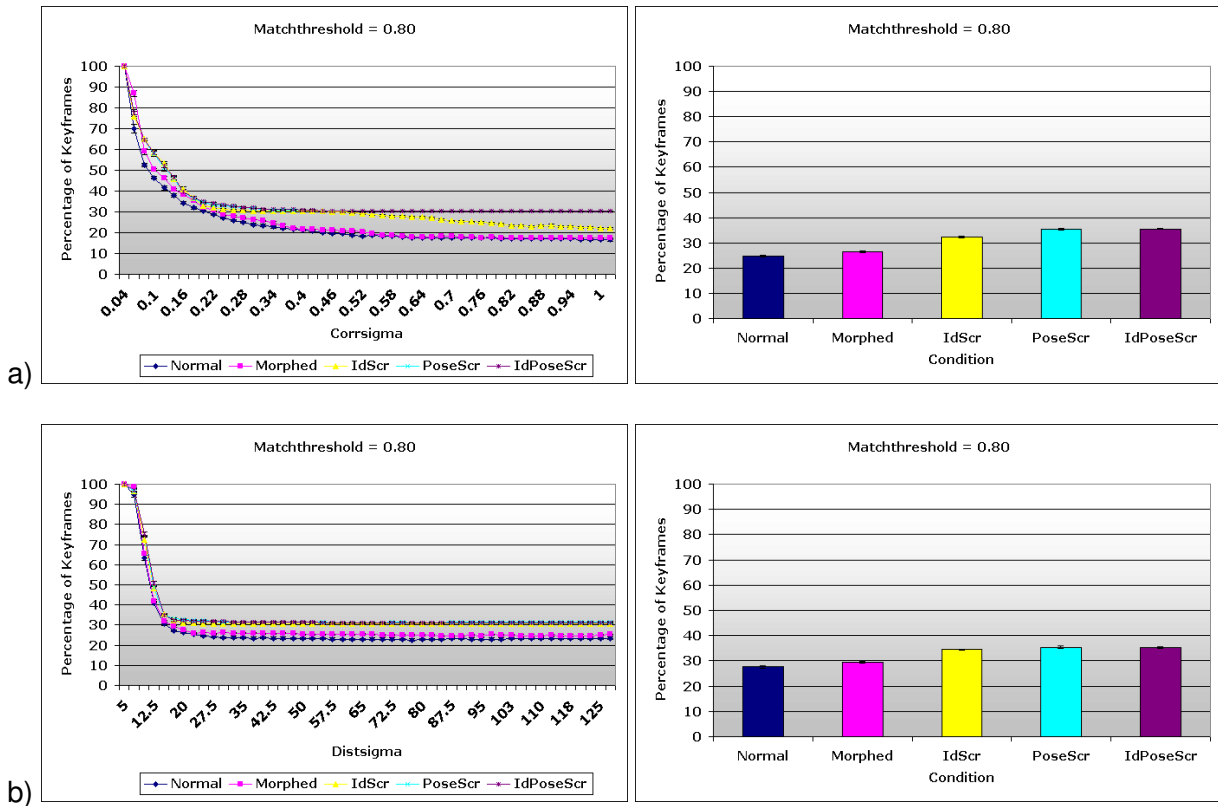
consistent advantage of the symmetrical $-15^\circ - 15^\circ$ sequence over each neighboring sequence suggesting the existence of a 'sweet spot' for which visual information of faces stays relatively constant during rotation.

In the context of the psychophysical results this first computational experiment has shown that tracking in the morphed condition comes, indeed, at the cost of a reduction in spatio-temporal coherence. Nevertheless, when compared to the results for the normal condition, the amount of trajectories extracted still allows to create a single coherent object representation. In addition, the linear variation of the visual complexity resulted in a linear extraction of trajectories demonstrating that the tracking framework delivers robust performance for both sparse and rich representations. Finally, the data for tracking across different viewpoints has shown some interesting results in terms of stable views for learning of visual information from faces. Although this is not immediately applicable to the present task of computational modeling of the temporal association experiments, the results nevertheless show how the tracking framework provides interesting and novel insights into the spatio-temporal processing of visual input.

4.3.3 Learning keyframes from morphed or scrambled sequences

The second set of experiments I want to discuss is based on the results of the previous experiment and is concerned with the build-up of keyframe representations from rotation sequences. Here, I want to trace the influence of the threshold parameters σ_v and σ_{app} that control the spatio-temporal coherence of the learning process.

Figure 4.14: Number of keyframes for normal, morphed, scrambled, pose-scrambled and fully scrambled sequences a) as a function of σ_a b) as a function of σ_v . The left plot shows the full curve, the right plot shows the averages of all five conditions.



4.3.3.1 Experiment 1 - Keyframes from morphed and scrambled sequences

Experimental Setup: For this experiment, sequences similar to the learning trials from the temporal association experiments were analyzed with the help of the keyframe framework. Stimuli (see Figure 4.13) consisted of 180° rotation sequences from 25 different faces in normal, morphed and three scrambled conditions. Each sequence contained 25 images from an equally spaced 7.5° sampling of the horizontal viewing axis. Normal and morphed conditions were created in the same fashion as in the previous experiment and are therefore similar to the experiment reported in Wallis [1998]. The first scrambled condition (IdScr) consisted of a consistent rotation with regard to pose (that is, from left to right profile) but with every fourth pose taken from a *different* individual. This condition was modeled after experiment 1 in the *second* study by Wallis [2002]. The second scrambled condition (PoseScr) contained 25 images from the *same* individual in which the sequence was divided into 8 blocks each of which showed a continuous rotation. These blocks were then scrambled in order to destroy any overall consistent rotation between two consecutive blocks. The third scrambled condition was modeled after experiment 2 in Wallis [2002] and contained sequences in which images were shown scrambled both with respect to pose *and* identity (IdPoseScr).

Each 180° sequence was presented to the system which extracted a number of keyframes. The percentage of keyframes (with respect to the original number of 25 images) was the crucial measure of this experiment. In addition to the five difference scrambling conditions, the influence of the parameters σ_v and σ_a on the keyframe extraction process was investigated (the remaining parameters, most notably $thresh_k$, were fixed). For this, one parameter was changed while the other was kept at the default value used in the previous section.

Results+Discussion: The results of this experiment are shown in Figure 4.14 which plots the dependence of the number and positions of the keyframes on the parameters σ_v and σ_a as well as the average number of keyframes (as a percentage related to the maximum number of keyframes possible) generated in the different conditions. In addition, Figure 4.15 shows the extracted keyframes in the five conditions.

Turning to Figure 4.14, the main conclusion that can be drawn from the plots is that after an initial drop in number of keyframes, all curves remain almost at a constant level for higher values of σ_v and σ_a . In addition, both normal and morphed condition show similar parameter dependencies that are significantly lower compared to the curves of the scrambled conditions (statistical tests on the average values confirm this to be highly significant with $p < 0.001$ for both normal and morphed conditions compared against each of the scrambled conditions). Even for rather "tolerant" similarity and speed values, all scrambled conditions consistently yield large numbers of keyframes. Taken as a measure of the temporal association, these results thus are consistent with the psychophysical experiments in which all scrambled conditions from the second study showed reduced effects compared to morphed sequences.

The difference between the parameter curves for the normal and morphed conditions is noticeable only for low values σ_a and higher values of σ_v with the normal condition resulting in less keyframes than the morphed condition. For the settings used in the previous experiments on the number of trajectories ($\sigma_a = 0.3, 0.3, 0.3$, $\sigma_v = 40, 20, 10$ for each of the three resolution levels), the normal and morphed conditions are clearly separated (see also Figure 4.15). This result is especially interesting as it provides a baseline for normal sequences against which morphed sequences can be compared. It seems to be the case that morphing does introduce small, but noticeable differences compared to normal rotation sequences which would indicate that temporal association should be easily possible in the keyframe framework. Confirmation for the ease with which morphed sequences are integrated comes from anecdotal evidence from experiments in which participants saw a rotating normal face sequence next to a rotating morphed sequence. When asked whether the two sequences depicted the same or a different face it took most participants several repetitions before any difference was reported. This demonstrates that even when direct visual comparison was possible it was hard to tell any difference between the two sequences - a finding that is reproduced in the results shown in Figure 4.14.

In addition, the averages show that pose (PoseScr) and identity (IdPoseScr) scrambling leads to a consistently higher strain on the keyframe learning process as identity scrambling (IdScr) only, which still preserves some form of spatio-temporal continuity. This result, however, is only found for the appearance term ($p < 0.001$ for σ_a). The velocity term seems not to be sensitive to the differences between the three scrambled conditions (all t-tests not significant for σ_v). If one takes the number of extracted keyframes as a measure for temporal association, this result reproduces the psychophysical findings from the second study, in which *no clear effect* was found for the fully scrambled condition as opposed to identity scrambling. In addition, it seems as if this is mainly the result of the appearance differences rather than the spatial differences between subsequent frames in the sequences.

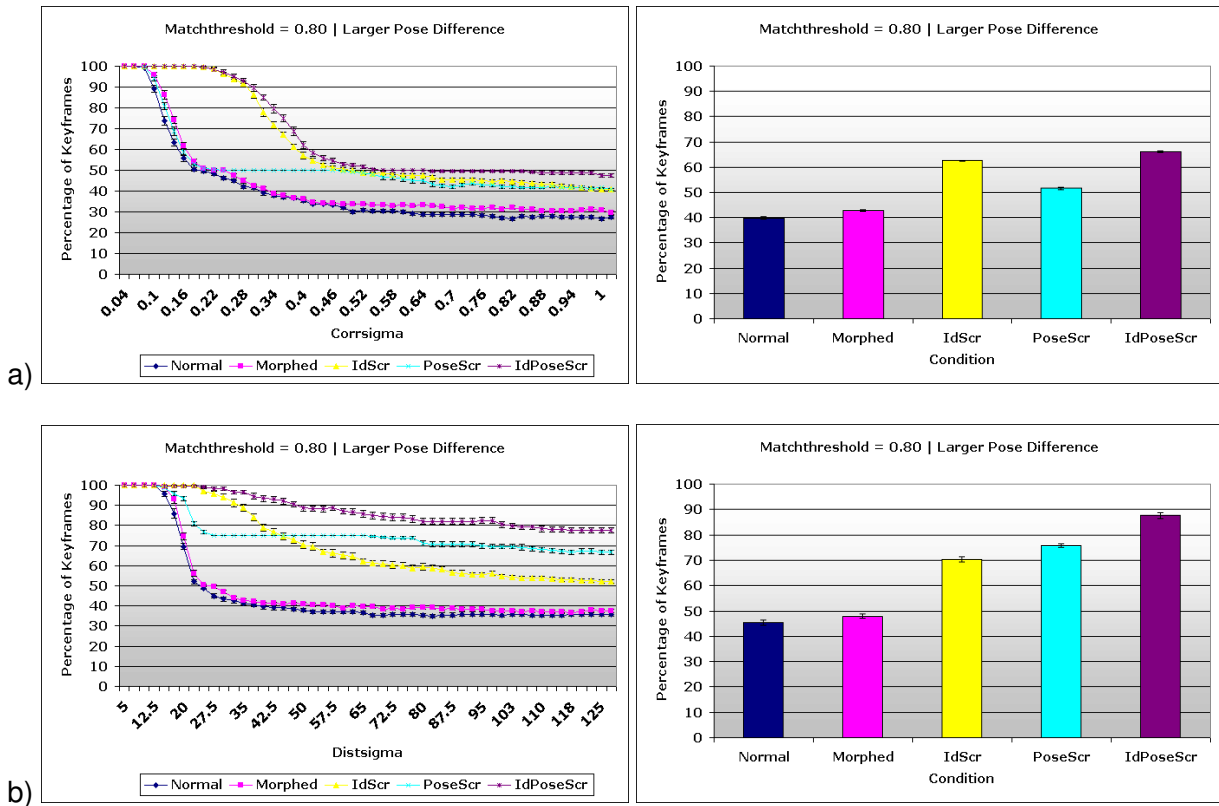
There are, however, two potential problems with this computational experiment, which I want to discuss. The first is that the visual difference between frames in the morph condition with respect to the normal condition was relatively small (for a depth rotation of 7.5° , the morph step was set to 15 percent). One possible reason for the small differences found in the experiment might thus be due to a low sensitivity of the chosen thresholds to the morphing manipulation. The second problem is that compared to the morph condition, visual differences between two consecutive frames in the scrambled conditions were relatively mild (scrambling only with respect to every fourth frame). Even though the experiment resulted in a relatively clear differential effect of appearance versus velocity threshold, it might have been the case that scrambling was too weak for the velocity threshold to show any effects.

In the following two experiments, I tried to address these questions by either

Figure 4.15: Extracted keyframes for the sequences from Figure 4.13 with feature trajectories for a) normal, b) morphed, c) identity scrambled (IdScr), d) pose scrambled (PoseScr) and e) fully scrambled (IdPoseScr) conditions. Note the different number of keyframes in the five conditions.



Figure 4.16: Number of keyframes for normal, morphed, scrambled, pose-scrambled and fully scrambled sequences for faster-changing sequences a) as a function of σ_a b) as a function of σ_v . The left plot shows the full curve, the right plot shows the averages of all five conditions.



- increasing the visual difference between two consecutive frames by increasing the pose difference by a factor of two (experiment 2)
- increasing the visual difference of the scrambled conditions with respect to the morphed condition by scrambling single images instead of blocks of images (experiment 3).

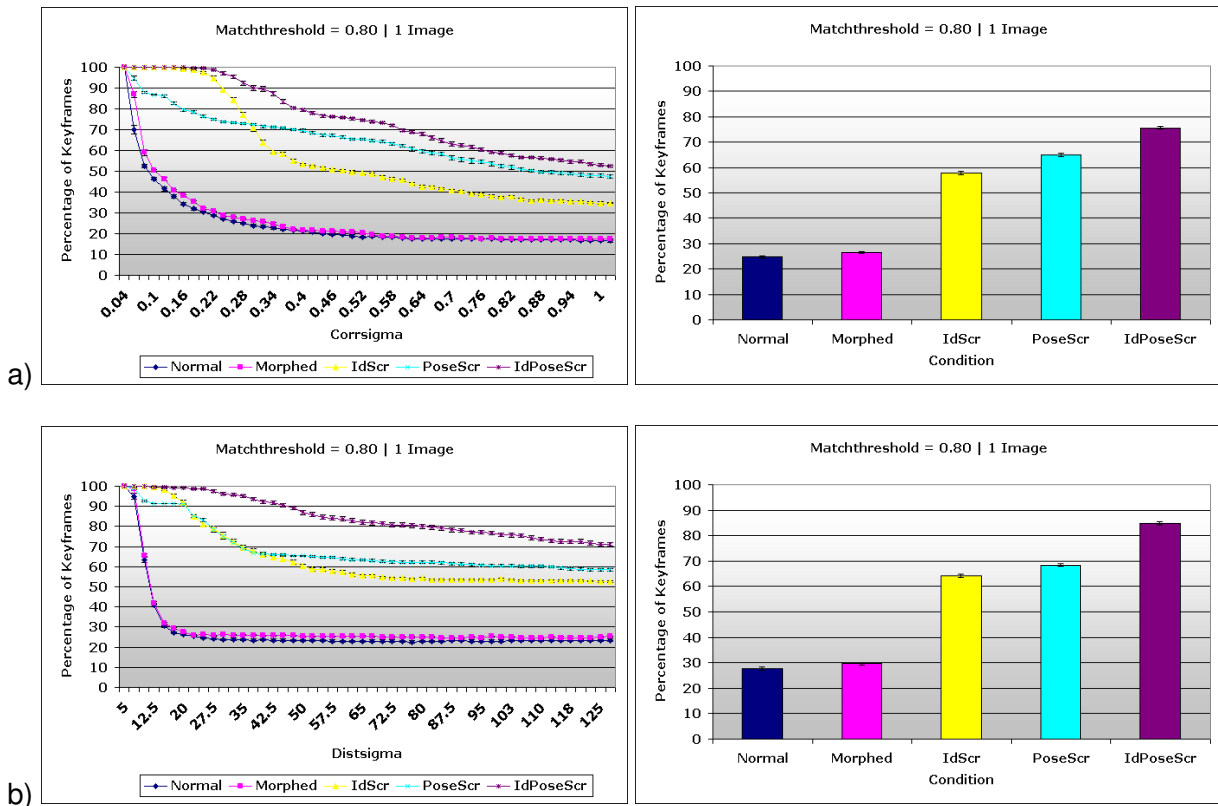
4.3.3.2 Experiment 2 - Keyframes from fast changing morphed and scrambled sequences

Experimental Setup: All sequences were the same as in the first experiment except that keyframe analysis was done with *half* of the images only. This resulted in a rotation sequence with an inter-frame pose difference of 15° and a total of 13 images for each sequence. All other algorithmic parameters were identical to experiment 1.

Results+Discussion: The results of this experiment are shown in Figure 4.16. Compared to experiment 1 the percentage of keyframes has increased for the normal and morphed condition indicating that the larger pose difference leads to an increased strain on the tracking process. In addition, the initial drop that was observed in the previous experiment, occurs over a larger range of parameter values after which the performance remains largely constant. The relative behavior, however, remains the same with respect to the normal and morphed conditions: both appearance and distance terms produce a small, but significant difference between the two conditions. This indicates that the previous results were *not* due to insensitive measures but confirms that any visual difference captured by the number of extracted keyframes between normal and morphed rotation sequences is relatively small.

Similarly to experiment 1, all scrambled conditions result in a significantly higher number of

Figure 4.17: Number of keyframes for normal, morphed, scrambled, pose-scrambled and fully scrambled sequences for completely randomized sequences a) as a function of σ_a b) as a function of σ_v . The left plot shows the full curve, the right plot shows the averages of all five conditions.



keyframes than either the normal or the morphed condition. In this experiment, however, the PoseScr condition actually yields less keyframes than the other two scrambled conditions when varying σ_{app} - the plot for σ_v even has a point where the curves from IdScr and PoseScr cross. On average, however, the spatial term in this experiment results in large differences between all three scrambled conditions. This stands in contrast to the previous experiment, where the conditions could not be separated with respect to the spatial term σ_v and shows that for the larger pose differences used in this experiment, there are relative changes in the feature distances that are diagnostic of the different types of scrambling.

In summary, the results support and extend the findings from experiment 1 by showing that both parameters are able to replicate the qualitative performance differences between the four conditions used in the psychophysical experiments.

4.3.3.3 Experiment 3 - Keyframes from morphed and fully scrambled sequences

Experimental Setup: Normal and morphed conditions were the same as in the first experiment. The first scrambled condition (IdScr) consisted of a consistent rotation with regard to pose (that is, from left to right profile) but with *every* pose taken from a different individual. The second scrambled condition (PoseScr) contained 25 images from the *same* individual which were then fully scrambled in order to destroy any overall consistent rotation between two consecutive images. Finally, the third scrambled condition was created by combining the manipulations from the previous two conditions.

Results+Discussion: The results of this experiment are shown in Figure 4.17. Compared to

Figure 4.14 it is immediately obvious that all scrambled conditions generate far more keyframes across all parameter values. Similarly to the previous experiment, the curves asymptote much later than in the first experiment. Both σ_v and σ_{app} clearly separate the three scrambled conditions in the averaged data. In addition, both plots have a cross-over point between the IdScr and PoseScr curves. Nevertheless, this third experiment also confirms the critical difference between the spatio-temporal similarity of the four psychophysical conditions.

4.3.3.4 General discussion

As I have already argued, the results from the psychophysical experiments show that for learning of objects, integration of images into one object representation is based on temporal contiguity, but with a strong influence of visual similarity between images. In this respect it might be better (as, indeed, the authors also did in their second article [Wallis, 2002]) to talk about *spatio-temporal continuity* instead. Such a learning process could then be modeled within the keyframe framework through feature tracking - the general assumption being that temporal association in the proposed framework can be linked to the number of generated keyframes. The important parameters determining the performance of the tracking framework are then given by appearance and distance/velocity terms that influence the similarity between consecutive frames.

The aim of the computational experiments was thus to investigate the influence of appearance and velocity thresholds on learning of keyframe representations. In general, one would expect differentiated results for all types of image manipulation which increase inter-frame visual similarity with regard to a normal (that is, unmanipulated) image sequence. Given that all manipulations in the psychophysical experiment necessarily included appearance changes as well as changes in feature distance (and thus velocity), one should expect both types of thresholds to show a differentiated behavior. Intuitively, the average number of keyframes should increase as follows: normal < morphed < identity scrambling < pose scrambling < pose+identity scrambling. This order is similar to the psychophysical findings in which temporal association effects were decreasing from morphed to identity scrambled to pose+identity scrambled sequences - it would be interesting to run the pose scrambling condition as well to see whether the results would indeed be between the other two conditions.

In general, the computational experiments have confirmed this effect throughout nearly all experimental conditions. For all experimental manipulations, a small but significant difference of the normal and morphed conditions was observed showing that morphed sequence presentation should result in a very similar visual representation of the spatio-temporal input. Given that the computational results aims at modeling temporal association based on the notion of "number of created keyframes" (one could also relate this to the concept of "visual complexity of representations"), one might ask how this would result in a differential recognition performance - especially since this parameter was not actually modelled in the experiments as the exact recognition task (recognition across 90° depth rotation) is still outside the scope of current face recognition systems. Within the keyframe framework, however, it is possible to provide two answers to the general problem of recognition, which I want to discuss in the following.

The first answer is based on the linked structure of keyframes: In the test phase of the experiment, a delayed match-to-sample task was done in which participants had to judge whether two sequentially presented images belonged to the same face or not. In the keyframe framework, such a task would amount to trying to match the first image to all keyframe representations in memory. As this particular pose was learned it is possible to retrieve a match from the database for this image. What happens to the second image? Recalling that keyframe representations consist of a *linked* graph of keyframes, once one of these frames is recognized, it can pre-activate (or "prime") the other keyframes in this representation. This seems to be a reasonable strategy for recognition of objects as it facilitates recognition of subsequent images of the same object. However, in the case of the psychophysical experiment it would result in priming the previously learned keyframes

from the *morphed* sequences. This in turn would influence the final recognition decision such that the number of false matches increased for the WS stimuli - as seen in the psychophysical experiments. From these arguments one can now define the recognition measure of temporal association based on the number of keyframes in the representation and the range of the activation within the representation (where range refers to the number of primed keyframes and with that also to the possible pose difference that can be pre-activated). If morphed sequences result in more keyframes, then recognition should be worse as a result of the decreased effective activation range. This would also explain why pose-scrambled sequences show virtually no effect as neighboring keyframes would not be able to prime neighboring poses effectively.

The second answer relies on the learning stage itself and is mainly applicable to the presentation of scrambled sequences: Recalling that each sequence is represented by a number of keyframes based on spatio-temporal continuity, it might also be possible that a scrambled sequence simply results in *multiple* object representations as the difference between consecutive images is too large for a single, coherent and consistent object representation. Evidence for this might come from the large number of keyframes compared to the morphed sequences. If object representations are broken apart, one should not expect any form of pre-activation to occur from one part to the next without a repeated training that would form links between representations.

Chapter 5

Computational studies I - Keyframes

After dealing with the psychophysical modeling capabilities of the proposed keyframe framework, this chapter deals with its computational validation. More specifically, the focus will be on computational studies on

- keyframe extraction from artificial and real-world sequences
- recognition of images with the help of several spatio-temporal keyframe representations
- incremental build-up of object representations from image sequences

The aim of this chapter is to show that the proposed framework is not only suited for modeling psychophysical findings but also - building on the first promising results in the previous chapter - for general purpose object recognition in a real-world context. Thus, experiments will be carried out on all aspects of the framework on both artificial and real-world image sequences.

5.1 Geometric constraints for local feature matching

As we have seen in the previous chapter in the context of psychophysical modeling, the use of geometric constraints results in increased recognition performance. In the following, I want to provide a more detailed exploration of the parameter dependencies of each of the constraints.

As in the previous chapter the task with which I want to assess the performance of the various feature matching strategies will be recognition under depth rotation. In particular, the following experiments will be conducted on the COIL database as well as on the face database. For an excellent in-depth analysis of local features focusing on appearance-based properties, see especially Mikolajczyk and Schmid [2005].

There is good psychophysical evidence that recognition of *faces* is possible even under viewing changes as large as 90° (see chapter 1 as well as Troje and Bühlhoff [1996]). In addition human recognition performance remains highly view-dependent across different viewing angles - a fact that seems to rule out complex, 3D analysis of faces as this would predict a largely view-invariant recognition performance (Biederman and Gerhardstein [1993]; but see also Biederman and Kalocsai [1997]). Such a 3D strategy in the form of morphable models (the system developed by Blanz et al. [2002], for example, achieves excellent performance), however, is currently one of the few methods, which are able to generalize across larger viewing angles (as well as illumination changes) from one image. So far, image-based methods - especially based on local features - have met with limited success in this task.

It is well-known that finding corresponding features between two images, which show the same face (or object) under a large rotation, is a very difficult problem (see Mikolajczyk and Schmid [2005] for a recent review of local feature descriptors). Apart from the fact that there is of course an upper bound to invariant extraction of features under large depth rotations, the issue of false

matches becomes more prominent as feature similarities start to decrease with increasing distance to the test view. In addition, the significant depth structure of faces makes the application of affine-invariant feature strategies especially difficult. Although it is difficult to find a large number of consistent matches across large pose differences, one type of information that can be used as an additional constraint is the configuration or geometric layout of the features. In the following, I will present computational experiments that explicitly test to which degree such simple, 2D information provides increased recognition performance across larger changes in viewing angle.

5.1.1 Geometric constraints

Following the terminology introduced in chapter 2 and chapter 3, geometric constraints will be considered in the local feature matching framework, where each feature carries appearance as well as geometric information. Features in two images are matched using the similarity matrix A , where each entry $A(i, j)$ is:

$$A(i, j) = \exp\left(-\frac{1}{2\sigma_{\text{geo}}^2} \text{geo}(\vec{p}_i, \vec{p}_j)\right) \cdot \exp\left(-\frac{1}{2\sigma_{\text{app}}^2} \text{NCC}(\vec{l}_i, \vec{l}_j)\right)$$

In the following, I will examine three different variants of this similarity measure which depend to different degrees on the 2D feature positions \vec{p} in the image.

- Standard: $\text{geo}(\vec{p}_i, \vec{p}_j) = 0$, where only appearance information is used during matching
- Euclidean distance: $\text{geo}(\vec{p}_i, \vec{p}_j) = \|\vec{p}_i - \vec{p}_j\|^2$, which is one of the most basic constraints penalizing large pixel distances between two features
- Embedding: $\text{geo}(\vec{p}_i, \vec{p}_j) = \sum_{k=1}^N \|d_i(k) - d_j(k)\|$, where \vec{d} is a sorted vector, which for any given feature contains Euclidean pixel distances to all other features in the image (see section 4.2)¹.

5.1.2 Recognition under large view rotations

Experiment Design: In the following, the three different local feature algorithms were benchmarked in a recognition experiment in order to evaluate their performance under large viewing changes. Two additional algorithms were chosen to provide a suitable baseline: the first algorithm simply consists of the standard Euclidean distance between images - a rather coarse, yet sometimes still surprisingly powerful algorithm [Sim et al., 2000]. The second matching algorithm is a state-of-the-art local feature framework based on scale-invariant features (SIFT, Lowe [2004]), which was shown to perform very well in a number of object recognition tasks. Local features in this framework consist of scale-invariant, high-dimensional (each feature vector has 128 dimensions) histograms of image gradients at local intensity maxima. The SIFT algorithm is available for download at <http://cs.spider.uk.ca/~lowe/> and was used without modification in the following experiments.

The databases used in the following experiments are the MPI face database as well as the COIL object database. From the face database, images were created in 5 different poses: -90° , -45° , 0° , 45° and 90° . Each grayscale image has a size of 256x256 pixels with the face rendered on black background. The whole set was split into a training and test set each containing 50 faces, where the training set consisted of the frontal (0°) and profile face views ($\pm 90^\circ$) and the test set of the intermediate ($\pm 45^\circ$) views. For the COIL database, the images subtended 128x128 pixels and were split into training and test sets each containing 50 objects with the training set consisting of the ($0^\circ, 90^\circ, 180^\circ$) views and the test set of the intermediate ($45^\circ, 135^\circ$) views. Similarly to the

¹Note that this list does not include *model-based* geometric features such as, for example, checks for affine transformations or rigidity assumptions.

psychophysical experiments as well as the computational experiments reported earlier, an old-new recognition task was chosen to benchmark the algorithms. In order to increase the statistical significance, the experiment was repeated 10 times with different training and test sets.

Results+Discussion: Standard matching: Figure 5.1a-c shows recognition performance as AUC values as a function of changing the size of the image, changing the matching threshold, as well as changing the number of features in the image. Not surprisingly, both increasing the number of features as well as increasing the matching threshold on average result in an increase in performance. The optimum performance as a function of image size, however, is attained at the lower end of the scale for faces and at the higher end for the COIL database. In addition, the relative performance increase that can be achieved is much higher by increasing the matching threshold and the number of features rather than by varying image size. Since the three geometric constraints use the same algorithmic structure, the optimal parameter values from this set of experiments will be used as a *baseline* to assess potential performance improvements in all subsequent experiments.

Matching with Distance and Embedding constraints: Figure 5.2a-b show recognition performance of the two geometric constraints when changing the relative influence of the constraint through adjusting $\sigma_{emb,dist}^2$ and changing the matching threshold. Increasing the influence of the embedding constraint results in an increased performance in particular for the face database whereas performance gains for the COIL database are marginal. In contrast, increasing the influence of the distance constraint results in a performance *decrease* for the face database as the feature distance for a 45° rotation becomes too large - this is less true for the COIL database where restricting the feature distances can help increase performance for some of the objects. In Figure 5.2c-d the influence of varying the matching threshold on both embedding and distance constraints is shown. Whereas a clear increase for both COIL and face database is seen for the embedding constraint with increasing matching threshold, for distance matching increasing the threshold increases discriminability only for the face database. This performance pattern for the COIL database is mostly due to the large appearance variation caused by the scale changes of the features between consecutive frames which therefore leads to better discriminability for lower matching thresholds.

Performance comparison: In Figure 5.3 the *best performance values* for all three local feature matching strategies were gathered and plotted for comparison against the baseline algorithms (Euclidean L2 matching and Lowe's local feature matching). First of all, it is easy to see that simple Euclidean image matching provides much worse performance than Lowe's local feature matching algorithm for both databases. In addition, all three local feature matching strategies outperform the two baseline algorithms for both databases. The advantage is particularly large for the face database which due to its controlled appearance seems to lend itself well to an analysis based on correlation of image patches. Both geometric constraints provide additional, significant increases in recognition performance for the face database. The best value for the embedding constraint shows marginally better performance than the best value for the distance constraint. The non-uniform view variation of the objects in the COIL database, however, results in a large decrease in performance for the embedding constraint.

In this experiment, I have investigated the recognition performance of two simple geometric constraints derived from a straightforward local feature algorithm under large depth rotations of 45° for two databases. For the MPI face database, excellent recognition results were obtained using the embedding constraint showing how perceptually motivated algorithms can improve recognition performance in a computer vision context. Owing to its non-regular image motion, however, optimum performance for the COIL database was achieved using the distance constraint. Both standard and geometric local feature matching methods in addition were found to provide a significant performance increase compared to either image-based matching or the standard implementation of the SIFT features. Although these results seem promising, the performance for the COIL database in particular needs to be improved. In the next section, I will therefore fo-

Figure 5.1: Average recognition performance for the standard matching algorithm for changing a) image size, b) matching threshold, c) number of features.

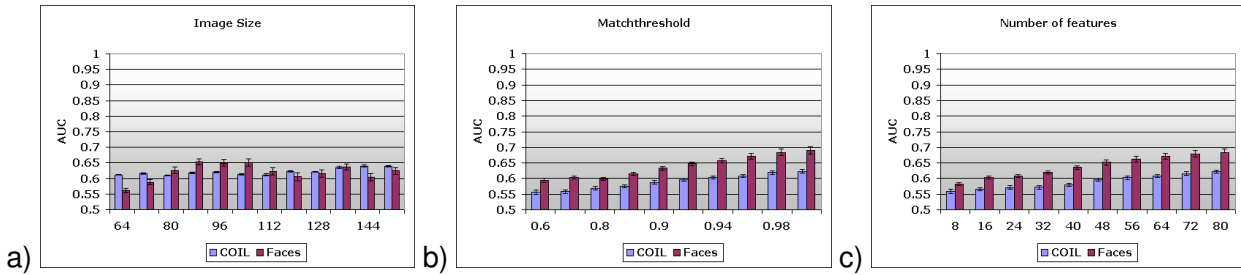


Figure 5.2: Average recognition performance for changing the relative influence of the constraint for a) embedding and b) distance, and changing the matching threshold for c) embedding and distance constraints.

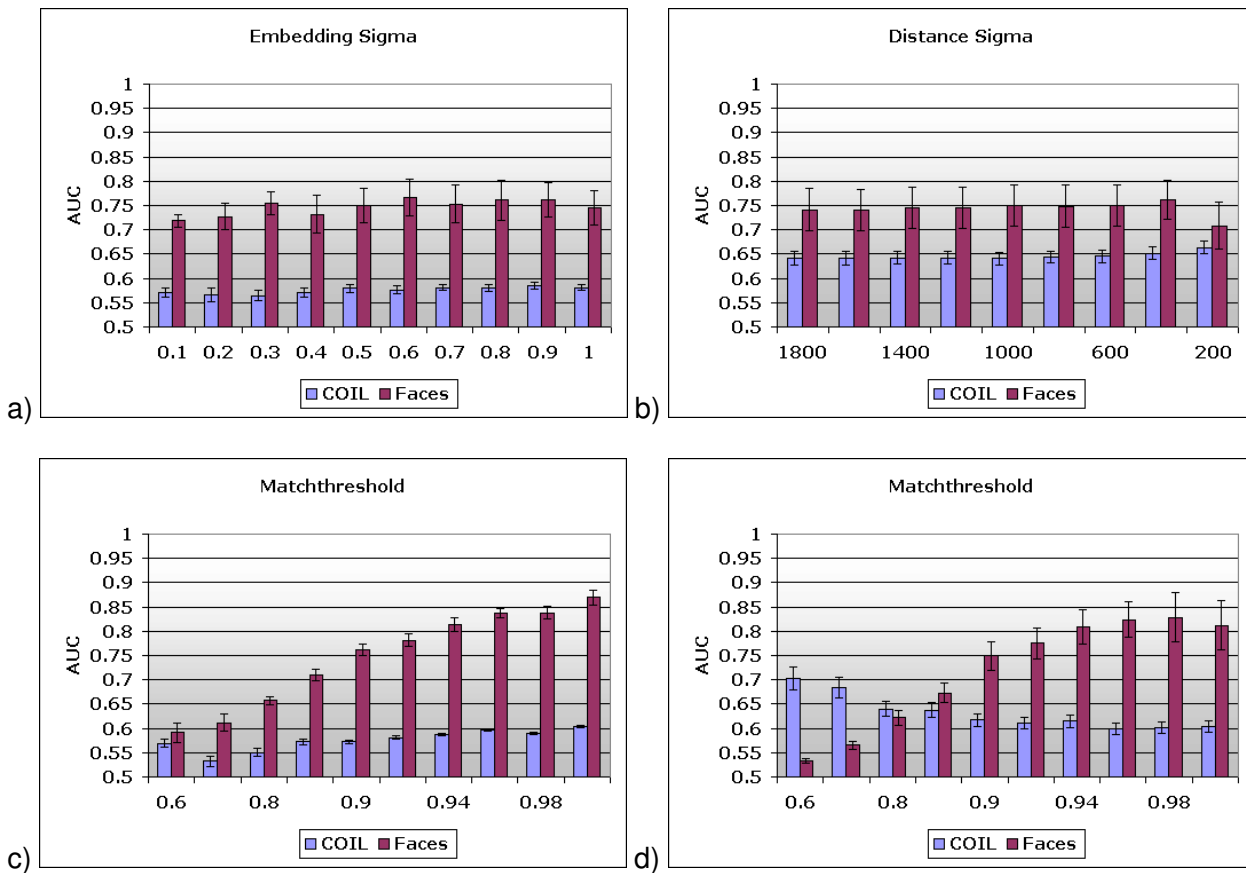
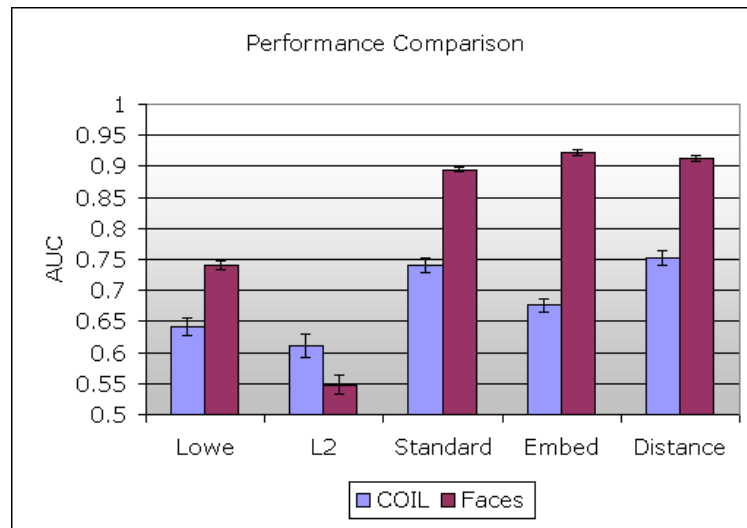


Figure 5.3: Recognition performance for three different local feature matching algorithms compared to L2 and Lowe’s local feature method as baseline.



cus on how keyframe extraction not only enables the learning of sparse, spatio-temporal object representations, but also on how recognition results can be further improved.

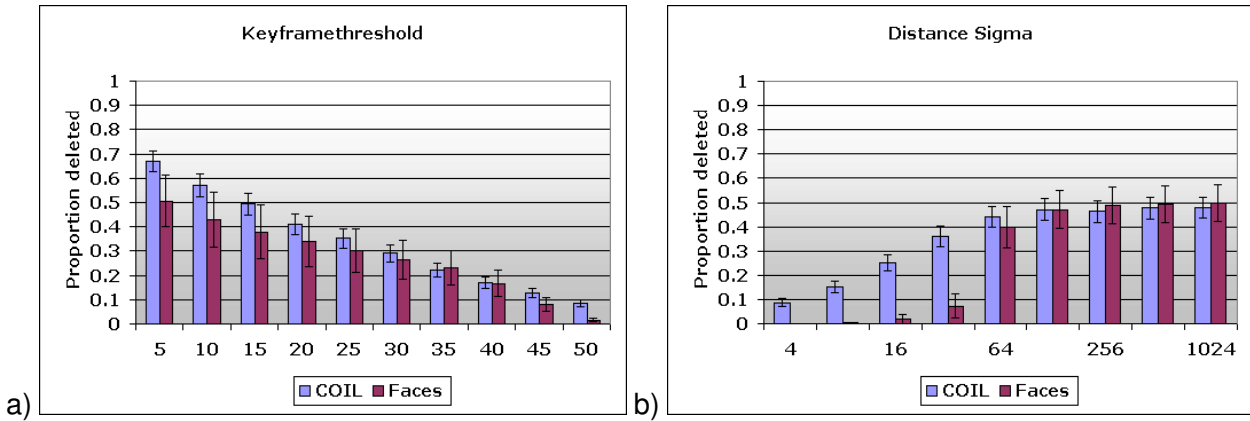
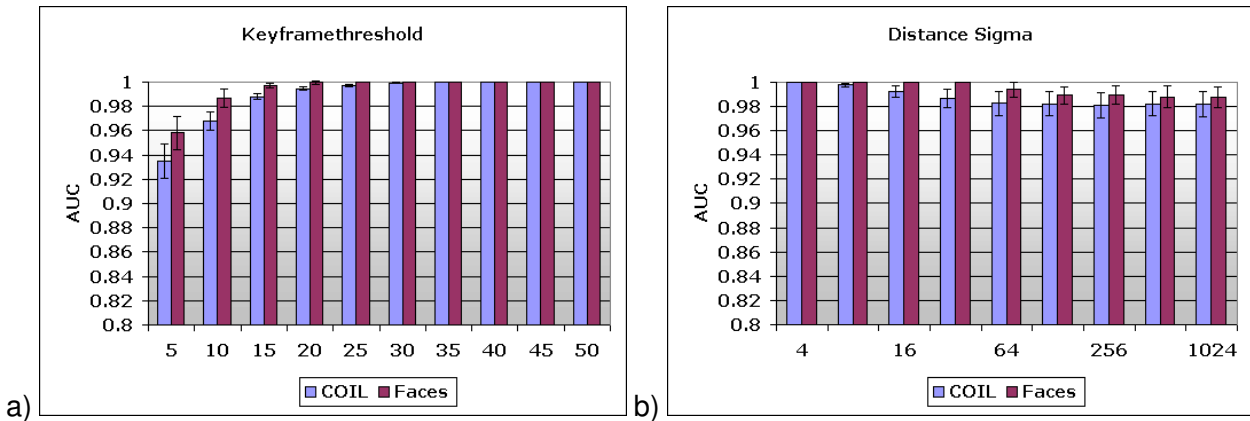
5.2 Keyframe extraction for learning of object representations

So far I have used the traditional computer vision paradigm for recognition: a pre-labeled database of static images for training and testing. Here, I want to show how by using the keyframe framework it becomes possible to learn view-based object representations similar to the ones used up to now from *dynamic* input data. Apart from providing a way of automatic learning of object representations, keyframes also enable the learning of geometric matching priors from the feature trajectories for increased recognition performance.

5.2.1 Parameters of keyframe extraction

Experiment Design: First of all, the keyframe extraction process will be investigated on both the face database and the COIL database. For the experiments the databases were randomly split into a target set and a distractor set, where each set contained 50 faces or objects. Keyframe extraction as detailed in chapter 3 was carried out on each face or object sequence in the training set using the standard tracking algorithm. This extraction process resulted in a set of keyframes and a set of deleted non-keyframes. Parameters that were varied for both databases were the keyframe threshold (that is, the percentage of features that will be tracked from one keyframe to the next) and σ_{dist} (that is, the average displacement of features *during tracking*). The set of keyframes was then used in an old-new paradigm (using standard local feature matching) with target images (all deleted images) and distractor images (images from other sequences of faces or objects) in order to determine the recognition performance of keyframe representations.

Results+Discussion: Figure 5.4a shows the results of varying the keyframe threshold for the COIL and face database, respectively. As can be expected, increasing the percentage of features that need to be tracked continuously from one frame to the next results in less deleted frames and thus in a less compact representation. This relationship is almost linear for both databases, which makes the keyframe threshold into an intuitive parameter that can control the complexity of

Figure 5.4: Proportion of deleted frames as a function of a) the keyframe threshold and b) σ_{dist} .Figure 5.5: Recognition results as a function of a) the keyframe threshold and b) σ_{dist} .

the resulting view representation. Interestingly, the COIL database on average generates a *lower* proportion of keyframes. Although the database has less regular image motion, some objects result in very stable tracking of features - this can be observed at high values of the keyframe threshold for which frames continue to be deleted from the COIL sequences.

The reason why the face database generates more keyframes is shown in Figure 5.4b, which plots the proportion of deleted frames as a function of the average pixel distance between consecutive feature points during tracking. In general, increasing σ_{dist} leads to better tracking of features across frames and therefore to fewer generated keyframes. Beyond a certain value of σ_{dist} , however, no improvement in tracking can be made and the proportion of deleted frames reaches asymptotes. Furthermore, this value is directly correlated with the average frame-to-frame feature distance in the sequences. As can be seen, the regular motion of the features in the face database results in an abrupt increase in deleted frames around a value of $\sigma_{dist} = 64$ after which saturation is quickly reached. In contrast, for the COIL database even at low values of σ_{dist} features continue to be tracked across frames, which in turn results in less keyframes per sequence on average.

Finally, in Figure 5.5 recognition results (again given as the area under the ROC curve (AUC); note also the different scale compared to previous figures) are plotted as a function of the keyframe threshold and σ_{dist} . Even for low values of the threshold recognition values are very high, showing that only few keyframes are needed to adequately represent the input sequences. In addition, both databases reach near-perfect recognition performance at *similar* values of the threshold. This

result demonstrates that the keyframe threshold can be used to control the resulting complexity of the keyframe representation independently of the underlying image data (at least for the two databases tested). In addition, changing σ_{dist} during tracking - while having a large impact on the number of generated keyframes - has only a minor influence on recognition performance.

One of the ways to determine an optimal combination of parameters is to take the set of parameters which results in the *least* number of keyframes while still guaranteeing perfect recognition. Figure 5.8a-b show the keyframes for the first 20 objects and the first 20 faces for each database for this set of parameters. As one can see, the number of keyframes varies depending on the depicted object for the COIL database (on average, 6.46 ± 2.58 keyframes were created): objects for which the bounding-box is resized (toy cars, boxes, etc.) generate a considerably larger amount of keyframes than objects with an upright symmetry axis (bottles, cans, etc.). In contrast, the regular motion in the face database results in only 4 keyframes across *all* sequences (see also chapter 4) - in addition, the viewpoints that are selected as keyframes are similar across sequences.

The overall advantage of the keyframe framework can be determined by comparing the recognition performance for object representations having the same average number of *regularly spaced* views. For the COIL database, this would amount to 6 or 7 frames per sequence and results in a recognition performance of **0.923 ± 0.015 (AUC - 6 frames)** or **0.943 ± 0.021 (AUC - 7 frames)** using standard local feature matching - considerably less than the perfect recognition score of **1.0 (AUC)** obtained with the keyframe views. Not surprisingly, for the face database with its evenly distributed keyframes, recognition performance decreases only slightly to **0.981 ± 0.011 (AUC)**.

5.2.2 Real-world sequences

Experiment Design: After presenting results on well-controlled image database, I want to show the feasibility of the keyframe framework on real-world data (see also chapter 8 for more real-world experiments). For this, we chose a database of car sequences, which was taken with a standard digital video camera. Each sequence showed a walk around a car - without taking control of sequence length, viewing condition, distance to the car, etc. As in the previous experiment, the database was randomly split into a target set and a distractor set, where each set contained 25 cars. Keyframe extraction was carried out on sequence in the training set using the standard tracking algorithm and with the optimal parameter set from the previous experiment. This extraction process resulted in a set of keyframes and a set of deleted non-keyframes. The set of keyframes was then used in an old-new recognition paradigm (using standard local feature matching) with target images (all deleted images from each sequence) and distractor images (images from other sequences of cars) in order to characterize the recognition performance of the resulting keyframe representation.

Results+Discussion: Figure 5.8c shows the resulting keyframes for the first 20 sequences. First, one can see that a similar amount of keyframes is generated for the same viewing range. The proportion of deleted frames for these sequences lies at **0.9** on average. Second, the performance in the old-new recognition task was at **1.0 (AUC)** - a further reduction in the keyframe threshold (data not shown) resulted in a reduced performance. This demonstrates the validity of the optimal parameters derived on the face and COIL database also for real-world data.

5.2.3 Recognition using keyframes

In addition to the automatic extraction of views, one of the advantages of keyframe extraction lies in the access to the *trajectories* of the tracked features. Provided the tracking itself is robust, this allows the extraction of detailed information about the 2D transformation that the local features have undergone between keyframes. In the following, I will discuss five computational approaches that make use of this information during recognition in different ways.

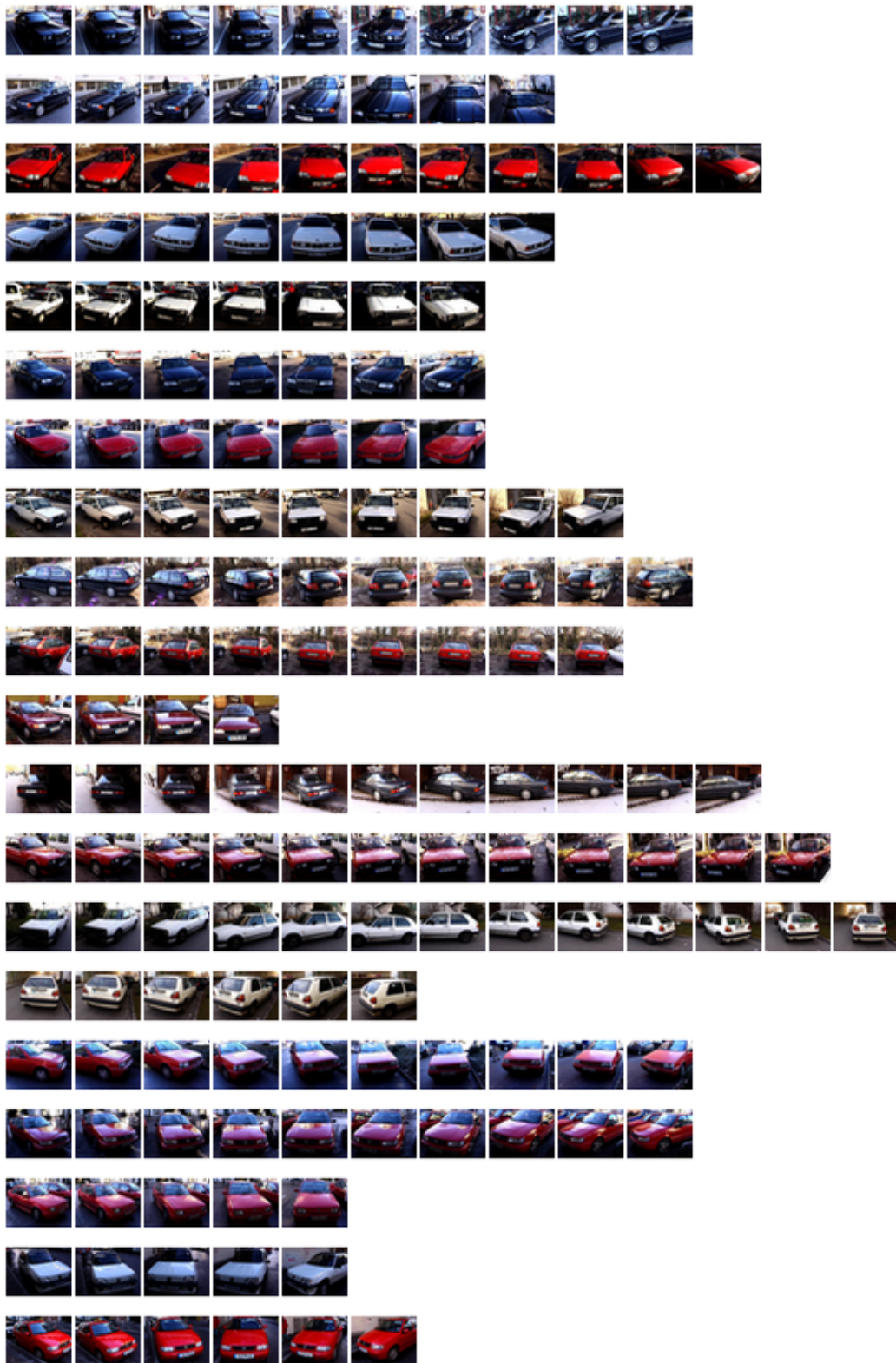
Figure 5.6: Keyframes for the first 20 objects of the COIL database.



Figure 5.7: Keyframes for the first 20 faces of the face database.



Figure 5.8: Keyframes for the first 20 cars of the car database.



All approaches rely on the basic matching equation introduced in chapter 3 using geometric matching, in which case the similarity matrix A for local feature matching becomes:

$$A(i, j) = \exp\left(-\frac{1}{2\sigma_{\text{geo}}^2} \text{geo}(\vec{p}_i, \vec{p}_j)\right) \cdot \exp\left(-\frac{1}{2\sigma_{\text{app}}^2} \text{NCC}(\vec{l}_i, \vec{l}_j)\right)$$

The first approach, feature trajectories were used to determine a matching prior (see chapter 4) for local feature matching. The geometric information is determined by the difference between the *average feature transformation* between two keyframes and the transformation between each feature pair i, j . That is, $\text{geo}_1(\vec{p}_i, \vec{p}_j) = \|\vec{n}_k - n_{i,j}\|^2$ where $\vec{n}_k = \text{normalize}(\frac{1}{n} \sum p_{k_1}^{\vec{p}} - p_{k_2}^{\vec{p}})$ if n features could be tracked between keyframe k_1 and k_2 and $n_{i,j} = \text{normalize}(p_i^{\vec{p}} - p_j^{\vec{p}})$. For each keyframe, this results in an additional two float values that need to be saved for encoding \vec{n}_k .

The obvious disadvantage of using average transformations during matching is that in order to be effective for a *global* matching prior the underlying image motion needs to be regular. One way to overcome this disadvantage is to make use of the *full information* provided by the feature trajectories as geometric information during local feature matching. This information can be used in the following manner: Given a feature trajectory T_i that belongs to a point \vec{p}_i in a keyframe and that consists of a set of t tracked points $\{p_{i_1}^{\vec{p}}, \dots, p_{i_t}^{\vec{p}}\}$, $p_{i_1}^{\vec{p}} = \vec{p}_i$ where t is equal for all T_i , each point $p_{i_t}^{\vec{p}}$ of a test image *votes* for an index t' in the trajectory iff $\|p_{i_t}^{\vec{p}} - p_j^{\vec{p}}\|^2 \leq \text{dist}$. The votes are collected for all points $p_{i_t}^{\vec{p}}$ of the test image and all trajectories T_i and the index t_{max} with the maximum number of votes is chosen as the most likely candidate for the geometric feature transformation occurring between the keyframe and the test image. The geometric similarity term then becomes:

$$\text{geo}_2(\vec{p}_i, \vec{p}_j) = \begin{cases} 1 & \text{iff } p_j^{\vec{p}} \text{ voted for } t_{\text{max}} \\ 0 & \text{otherwise} \end{cases}$$

This, however, results in feature matching that solely relies on similarity between *tracked* features and therefore needs a minimum number of trajectories in order to provide discriminability. Another possibility is therefore to integrate contribution from *other* local features as well, in which case the geometric similarity term becomes:

$$\text{geo}_3(\vec{p}_i, \vec{p}_j) = \begin{cases} 1 & \text{iff } p_j^{\vec{p}} \text{ voted for } t_{\text{max}} \\ \|\vec{p}_i - \vec{p}_j\|^2 & \text{otherwise} \end{cases}$$

In the following experiment, I will test five different recognition algorithms:

1. Recognition using $\text{geo}_1(\vec{p}_i, \vec{p}_j)$, that is, average transformations including appearance information (geo1)
2. Recognition using $\text{geo}_2(\vec{p}_i, \vec{p}_j)$ in conjunction with appearance-based information (geo2/app)
3. Recognition using $\text{geo}_2(\vec{p}_i, \vec{p}_j)$ without appearance-based information thus relying solely on geometric feature information (geo2)
4. Recognition using $\text{geo}_3(\vec{p}_i, \vec{p}_j)$ in conjunction with appearance-based information (geo3/app)
5. Recognition using $\text{geo}_3(\vec{p}_i, \vec{p}_j)$ without appearance-based information thus relying solely on geometric feature information (geo3)²

Experiment Design: Recognition was investigated on the keyframe representations of the face database and the COIL database. In order to demonstrate possible improvements due to the geometric matching prior, the keyframe extraction parameters were chosen as the ones delivering the *worst* recognition performance (see previous section). Again, the set of keyframes was used

²Note that geo2 uses only information from the tracked features, whereas geo3 uses information from *all local features* in a keyframe.

in an old-new recognition paradigm with target images (all deleted images from each sequence) and distractor images (images from other sequences of faces or objects) in order to determine recognition performance.

Results+Discussion: Figure 5.9 shows the results for the five different matching strategies in comparison to the best performance obtained with standard local feature matching. Before going into detail, it is worth mentioning that *all five* matching strategies result in a significant increase in performance, which demonstrates the benefit of incorporating information about tracked features into the matching process.

The average feature transformation between two keyframes (geo1) seems well suited for the very regular feature motion of the face database resulting in near perfect performance. For the COIL database, however, the fact that the sometimes rather complex image trajectories are approximated by an average transformation results in a smaller increase in performance. This increase is mainly due to the objects that have a more regular feature motion in the image plane.

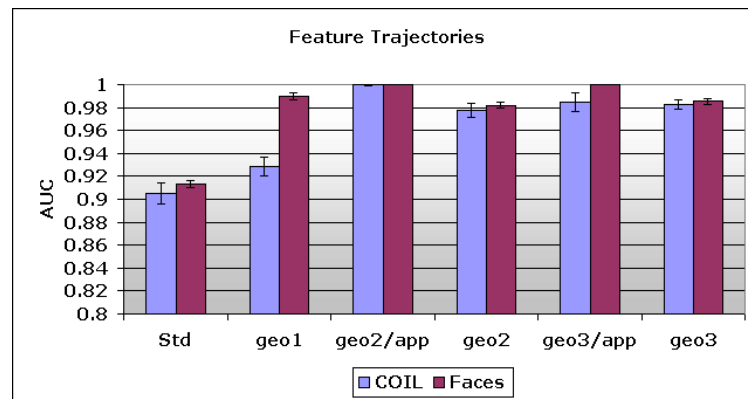
Performance using a combination of geo2 and appearance-based information is perfect for both COIL and face database. For the face database, there were on average 7 feature trajectories per keyframe at the specified keyframe threshold of 5%, resulting in a total of 144 keyframes for all 50 sequences - this is equivalent to only 2.9 keyframes for a full 180° rotation of a face. Storing the appearance-based information in each keyframe as 5x5 pixel image fragments therefore requires $144 \cdot 7 \cdot 25 = 24.6$ kByte. As each of the original sequences had 13 frames, this results in an average trajectory length of 4.5 frames and a total of ~4500 trajectory points for all sequences. For two values per trajectory point this would amount to an additional 8.8 kByte, which results in a **total storage requirement of only 33.4 kByte for all 50 sequences while ensuring perfect recognition**. Using geo2 without appearance-based information, however, recognition performance is still excellent at around 0.98 (AUC), which demonstrates the discriminability of the pixel positions that are contained in the feature trajectories. The resulting keyframe representation would for the face database only need to consist of the *7 feature trajectories for each keyframe*, that is, **8.8 kByte** - a **compression factor of ~2900** compared to the original sequence size!

Finally, using geo3 and geo3/app, the same performance pattern can be observed for the face database, whereas for the COIL database geo3/app does not result in an increase in performance compared to geo3. Although by including the appearance term for all features (as opposed to geo2/app, which only includes the appearance term for the tracked features) the amount of correct matches increases, the amount of *false matches* also increases for some of the objects in the COIL database. For the face database, storage requirements for this keyframe representation are increased, as all non-tracked features also need to be stored - in this case, the whole representation needs $144 \cdot 120 \cdot 25 = 420$ kByte for the image fragments and $144 \cdot 120 \cdot 2 + 4500 \cdot 2 = 42$ kByte for feature points as well as feature trajectories, resulting in a total of **462 kByte for all 50 sequences**. Nevertheless, this still corresponds to a **compression factor of 55** compared to the original sequence size.

In summary, including feature trajectory information results in significant performance improvements for both databases. Best performance can be obtained using a very sparse keyframe representation which consists of tracked feature points in combination with image fragments. Although this representation ensures high discriminability, retaining only the tracked features would only be optimal if the task were to recognize images from the originally seen sequences. In order to achieve additional robustness, one could use rotation- and scale-invariant matching to collect the votes for the trajectory points, which would help to recognize rotated and scaled images. Nevertheless, to ensure better generalizability across different viewing conditions, it will be beneficial to retain the appearance information of *all* features (corresponding to the geo3/app condition).

One of the biggest criticisms one might raise against these recognition results is that they were obtained by using the *test* images - after all, the keyframes were extracted using feature tracking on the original sequences. Although this is true, I want to mention two points which counter this argument:

Figure 5.9: Recognition using feature trajectories. Shown are the baseline performance (Std) as well as the five different geometric matching strategies (see text).



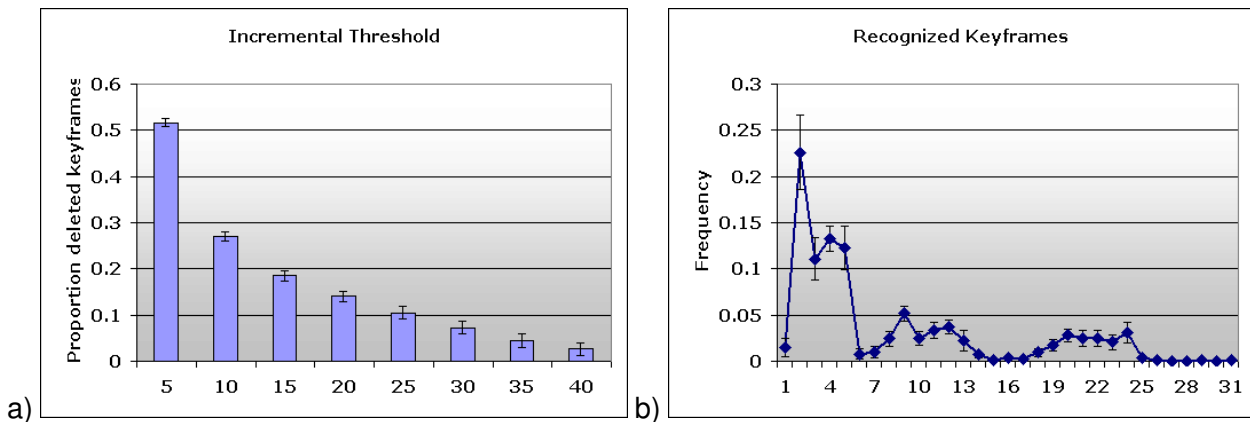
- One can see these experiments as a validation of the extreme sparseness that is needed in the keyframe framework in order to represent a given sequence. It becomes possible to ensure perfect recognition performance with less than 3 keyframes for both the face and COIL database.
- As mentioned above, it is possible to include both rotation and scale-invariance into the matching procedure in order to make it more robust to other viewing conditions.
- As was discussed in earlier chapters, humans learn in a spatio-temporal context and seem to generate inherently spatio-temporal object representations. Thus it makes sense also for computational systems to include spatio-temporal information. The relatively low storage impact of storing this information - ranging from 2Bytes/keyframe for geo1 to 32Bytes/keyframe for geo2, geo3 - makes this approach feasible even for *many different* sequences per object. The totality of keyframes in combination with feature transformations will ensure recognition across a wide variety of viewing conditions. This would also tie in to the "exposure" argument discussed in the experiments in chapter 1.2 as recognition performance would grow with exposure to the environment.

5.3 Incremental build-up of object representations

One of the properties of the proposed framework is the capability to incrementally learn object representations. In particular, as discussed in chapter 3, keyframes form a linked graph of views, which can easily be extended with new keyframes provided one of the new keyframes links to a previously learned keyframe. For this, keyframes are extracted in an image sequence as before. Each newly incoming keyframe then is compared to the already learned frames with local feature matching using a suitable recognition threshold³. If no match is found, the frame is added to the representation, whereas for each match a hitcounter for the corresponding keyframe is increased. The resulting representation is a linked graph of characteristic views of the presented sequences with additional information about the frequency of each view. This additional information can, for example, be used to speed up matching since it is reasonable to assume that new keyframes will match a highly frequented view (see chapter 1 for a discussion of the "canonical view" concept

³To ensure consistency, this recognition threshold should be equal to the standard recognition threshold for matching of images.

Figure 5.10: Incremental learning performance: a) proportion of deleted keyframes as a function of the incremental learning threshold (in percent), b) frequency histogram showing which keyframes were recognized during incremental learning.



to which this is related). In the following, I will present results from experiments on incremental learning that were conducted with the MPI face database.

5.3.1 Parameters of incremental learning

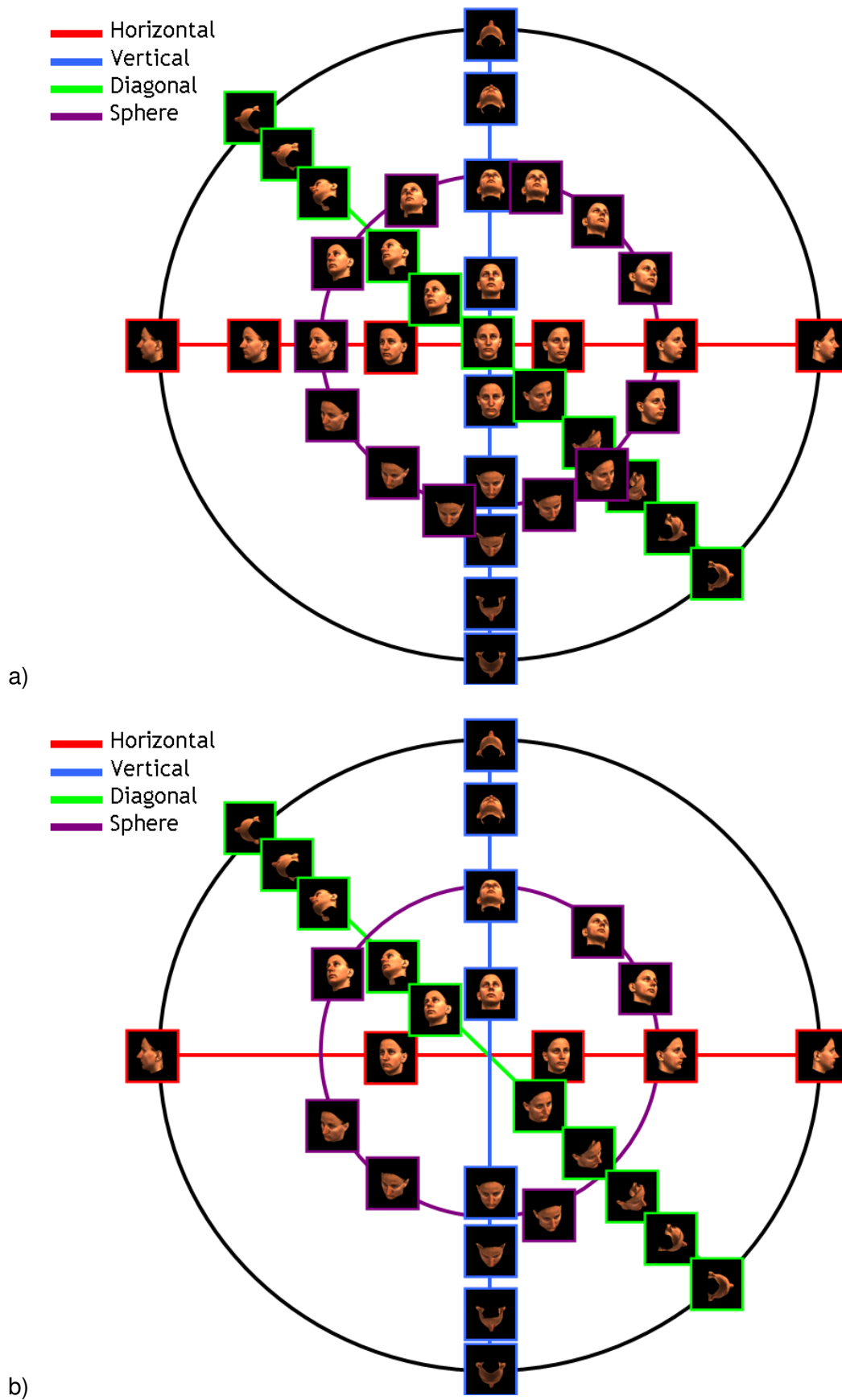
Experiment Design: For this experiment, 30 sets of sequences of rotating faces were created where each set contained four different pose animations. The first sequence consisted of the left-to-right rotation used in the previous experiment. The second sequence ran along the vertical viewing axis, the third sequence traced the viewing sphere diagonally from top left to bottom right, and the fourth sequence went in a circle around the view sphere (see Figure 5.11). For all sequences, keyframes were extracted using parameters derived from the keyframe experiments discussed in section 5.2.1.

Results + Discussion: Figure 5.11a shows keyframes for the four sequences of one face if they were treated independently with each keyframe at its corresponding position. For the set of parameters chosen, **38 keyframes** were extracted. Figure 5.11b shows the connected keyframe representation in its final state *after* incremental learning - only **29 keyframes** were retained, corresponding to a decrease of **24 percent**. As can be seen in comparison to Figure 5.11a, most keyframes were rejected for the final ("Sphere") sequence as the view space was already well covered by that point.

Figure 5.10a shows the proportion of deleted *keyframes* averaged over 30 faces as a function of the incremental recognition threshold, which specifies the match strength that a frame has to exceed in order to be rejected during learning. As could be expected, with increasing threshold less and less frames are rejected. Nevertheless, even for standard recognition thresholds of 10-15 percent match strength, **20-30%** of the keyframes are pruned. In addition, recognition performance for the resulting keyframe representation in the standard old-new recognition task remains at **1.0 (AUC)**, which demonstrates that pruning did not compromise generalizability.

Figure 5.10b shows a normalized frequency histogram plotting the average hits for each of the keyframes that make up the final object representation. Given the particular order of the sequences, the keyframes which received the most hits lie on the horizontal axis (with the second and third frame receiving the most hits). These keyframes would make suitable candidates for "canonical views". Finally, the speed-up that was achieved by matching according to previous viewing statistics were **15%** - a significant speed-up given the small size of the image sequences.

Figure 5.11: Keyframes for four sequences of a rotating face. a) all keyframes, b) keyframes after incremental learning.



5.4 Conclusion

The proposed framework is purely exemplar-based, that is, it does not rely on pre-constructed models or high-level a-priori knowledge. The main arguments against exemplar-based methods, which the keyframe framework addresses, are that:

- they require a large number of training examples,
- they use large amounts of storage for the representations,
- indexing into the representation takes a long time.

While the amount of training examples in our framework is still higher to obtain invariant recognition than with an underlying generative model (for example, Blanz et al. [2002]), the combination of geometric information from tracking with image-based, appearance information in the local feature matching process has shown excellent performance in the tested recognition tasks. This property thus enables the system to generalize over more cases in comparison to other image-based techniques using whole images, which in turn results in fewer training examples.

In addition, the proposed framework creates sparse representations which nevertheless retain some appearance information. This represents a compromise between the two extremes of purely image-based approaches and highly abstracted model-based approaches and allows while still allowing for good compression.

With regard to matching times for recognition, the complexity of matching is linear with the amount of trained sequences, linear with the amount of key-frames found and quadratic in the amount of features in each keyframe. The incremental learning technique described in the previous section, however, allows for a reduction in both training and recognition time since at some point the whole image space will be covered with keyframes. In addition, both the concept of the "canonical view" as well as the linked graph structure reduce recognition time by taking into account the viewing statistics of the learned representation.

As was stressed earlier, the aim of the computational implementation of the keyframe framework was less to provide performance superior to several state-of-the-art algorithms (although the recognition results presented in this chapter demonstrated surprisingly good recognition performance for the tested databases) but rather to demonstrate how spatio-temporal processing using local features results in efficient view-based object representations. It is possible to further improve the performance of the system by using

- different local feature descriptors: The keyframe framework is compatible with any feature descriptor - in particular, it would be interesting to include affine invariant feature descriptors such as listed in Mikolajczyk and Schmid [2005] which offer more robust matching performance than the image fragments used here
- multiple feature descriptors: So far, only one type of feature descriptor was used - given that human vision relies heavily on multiple cues, however, integrating several feature descriptors in a cue integration scheme would offer a further increase in generalizability
- more efficient tracking algorithms: For simplicity, tracking and recognition in the experiments presented in this thesis have used the same matching algorithm - although this results in a parsimonious implementation, more advanced tracking algorithms will be able to further improve performance by, for example, allowing more feature trajectories to be extracted
- feature transformations across multiple objects: The extracted feature transformations might be characteristic not only of the specific sequence but also of object classes or events - by analyzing feature transformations across several thousand sequences, such regularities could be extracted and be used as high-level priors for feature matching.

Chapter 6

Computational studies II - SVMs and local features¹

This chapter as well as the following deal with another key component of the object recognition process, namely the classification process. So far, classification, that is, the assignment of a label from a database of learned images to a novel image, has been dealt with in a fairly loose manner. Here, I want to approach the topic from a machine learning perspective.

Based on the distinction outlined in chapter 2, recognition performance in a computation recognition system is dependent on the choice of a suitable data representation and on the choice of a learning or classification algorithm which processes such representations. Motivated both by psychophysical and computational considerations on the efficiency and robustness of data representations the keyframe framework presented in chapter 3 relies on local appearance-based features. In addition to the data representation, chapter 2 presented a brief introduction into support vector machines - an algorithm from statistical machine learning, which has gained widespread attention in the field due to its firm theoretical grounding and outstanding generalization capabilities. In this chapter, I will present a framework which combines the two approaches in order to build a complete recognition system.

6.1 Introduction

Recognition of objects in unconstrained environments is one of the holy grails of computer vision since the advent of the field. Numerous approaches to this problem have been proposed which - according to the terminology of object recognition proposed in chapters 1 and 2 - can be broadly classified as belonging to two main categories: structural and appearance-based approaches. Structural approaches usually try to solve the recognition problem by explicitly reconstructing the 3D environment of the recognizing agent - recognition amounts to reconstruction. Appearance-based approaches on the other hand tend to focus on statistical descriptions of the image input data - recognition based on pixels.

In addition to the cognitive perspective, however, it seems that "pixels" have gained an advantage over reconstruction - at least in terms of progress on the previously stated problem of recognition and categorization of objects in real-world settings. As mentioned in chapter 2, within the field of appearance-based approaches there has been a gradual development of image representations starting from *global* representations (such as the simple, but surprisingly powerful pixel histograms) towards *local* feature descriptors, which are invariant to a range of image transformations (such as affine-invariant features). Local features have inherent properties, which make them ideal for recognition in real-world scenarios containing cluttered environments, occlusion,

¹This chapter is based on Wallraven et al. [2003].

changes in lighting, etc. Furthermore, the concept of local processing as very much consistent with the results from psychophysical and physiological studies (see chapter 3).

In the context of the framework introduced in chapter 3, a second important element to recognition is the process of learning and classification, which contains both building the training database from given examples and assigning labels to novel examples during testing. One of the recent developments in this area of classification schemes has come from statistical learning theory with the support vector machine (SVM) framework. This framework provides solid and mathematically proved optimality constraints for learning and classification problems and has received widespread attention in the literature. This is not only due to theoretical considerations but is also exemplified by an increasing number of computer vision work, which has proved that SVMs offer superior generalization capabilities and flexibility in a variety of computer vision tasks.

Recently, SVMs and kernel methods have begun to be used in combination with appearance-based object representations in the computer vision community. In a landmark paper, Pontil and Verri [1998] demonstrated the robustness of SVMs to noise, bias in the registration and moderate amount of partial occlusions. These results were obtained using raw pixel images as the most basic global object representations. Roobaert et al. [2001] examined the generalization capability of SVMs, when just a few number of views per objects are available. Barla et al. [2002] proposed to use a new class of kernels, which were especially designed for computer vision and inspired by similarity measures successfully employed in other vision applications (including histogram intersection and Hausdorff kernels).

Given these two recent developments in computer vision and machine learning the main contribution of this chapter is to show for the first time that they can be *fused* in a *local SVM framework*, which combines the best of both worlds - a robust representation and a state-of-the-art classification scheme. The remainder of the chapter is organized as follows: after introducing the general problem of combining local features and SVMs, a generic framework for local SVMs is proposed along with a demonstration of how several state-of-the-art local feature techniques can be integrated into this framework. Section 6.4 then reports extensive recognition experiments on a number of databases with noise-free and noisy data which show that the proposed framework yields superior performance when compared to other state-of-the-art approaches.

6.2 Support Vector Machines and local features

Given a set of images $\mathcal{I} = \{\vec{I}_i\}_{i=1}^m$, the most general representation of an image \vec{I}_i as a local feature vector can be described as $F_i = \{\vec{p}_j(\vec{I}_i), \vec{f}_j(\vec{I}_i)\}_{j=1}^{n_i}$, which is computed as follows (see also chapter 3):

- First, an interest point detector (a popular choice is the Harris corner detector) detects n_i points. An important point is that, in general, the number of interest points detected for each image \vec{I}_i will differ.
- The interest point detector yields pixel coordinates in the image plane in the form of $\vec{p}_j(\vec{I}_i)$;
- Finally, a feature descriptor is used to determine $\vec{f}_j(\vec{I}_i)$, which is a feature vector computed locally around the j -th point and for example might characterize the pixel variation in the neighborhood by means of a suitable descriptor.

When one does not consider interest point coordinates, the local feature vector reduces to $F_i = \{\vec{f}_j(\vec{I}_i)\}_{j=1}^{n_i}$.

Recalling from chapter 2 that the optimal separating hyperplane for SVMs is specified by

$$f(\vec{x}) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i K(\vec{x}_i, \vec{x}) + b \right) \quad (6.1)$$

it becomes clear that local features cannot be used in a straightforward way as an input. For this, there are mainly two reasons that are derived from computational properties of almost all local feature extraction algorithms and that I want to discuss in more detail in the following.

The first property of local feature extraction algorithms is that usually n - the number of features - is different for each image as it cannot be guaranteed that all images contain the same number of "interesting" local features. Whereas it is of course possible to impose an upper (or lower) limit on the number of features n , in general this is not advantageous as it severely restricts the robustness of the local feature representation. Consider for example a simple case in which the number of features is limited to n_1 and in which during training features are extracted from images of objects on a uniform background. If then during testing a novel image shows one of the previously trained objects on a *cluttered* background this will usually result in a much larger number of features $n_2 > n_1$ due to the increased pixel variation in the background. As $n_2 > n_1$, the scalar product in equation 6.1 is ill-defined and thus even a standard computer vision scenario - recognition in cluttered background - does not seem possible within the SVM framework.

One might argue that this could be avoided by adding an appropriate number of zeros to each feature vector in order to normalize vector lengths. This proposition can be examined by considering two local feature vectors $F_1 = \{\vec{f}_j(\vec{I}_1)\}_{j=1}^{n_1}$ and $F_2 = \{\vec{f}_j(\vec{I}_2)\}_{j=1}^{n_2}$, with $n_2 > n_1$ (the argument can be extended easily to the case of local features including point coordinates). We can now define a new feature vector \tilde{F}_1 by zero-padding F_1 :

$$\tilde{F}_1 = \underbrace{\{\vec{f}_1(\vec{I}_1), \dots, \vec{f}_{n_1}(\vec{I}_2)\}}_{n_1}, \underbrace{\{0 \dots 0\}}_{n_2 - n_1}$$

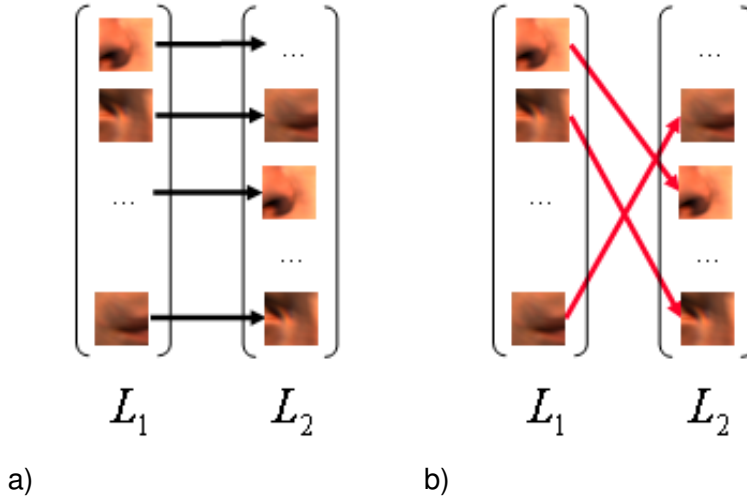
This definition would allow to compute the scalar product between F_2 and \tilde{F}_1 :

$$\begin{aligned} \tilde{F}_1 \cdot F_2 &= \vec{f}_1(\vec{I}_1) \cdot \vec{f}_1(\vec{I}_2) + \vec{f}_2(\vec{I}_1) \cdot \vec{f}_2(\vec{I}_2) + \dots + \\ &\vec{f}_{n_1}(\vec{I}_1) \cdot \vec{f}_{n_1}(\vec{I}_2) + 0 \cdot \vec{f}_{n_1+1}(\vec{I}_2) + \dots + 0 \cdot \vec{f}_{n_2}(\vec{I}_2) \end{aligned}$$

Whereas it is now technically possible to compute scalar products for local features, the computed quantity is not suitable from the point of view of recognition. The underlying philosophy in describing an image by local features is that once "interesting points" in the image are detected, local descriptors are computed around these points. Such a local descriptor should be *discriminative* in the sense that, if the point is detected again in a new image, the comparison of the descriptors computed around the points will allow them to match correctly. Thus one can see that local features are effective for recognition if and only if the algorithm we use measures similarities between *all* local features within the compared images. This is exactly what state-of-the-art algorithms for matching and recognition do (see, for example, Schmid and Mohr [1997], Schaffalitzky and Zissermann [2001], Laptev and Lindeberg [2003], Mikolajczyk and Schmid [2005], Lowe [2004]).

The second - very much related - property of local feature extraction algorithms which make a straightforward integration of local features into the SVM framework problematic is given by the correspondence problem. To illustrate this, let us assume that the lengths of all feature vectors n are the same, which obviates the first problem and guarantees that the scalar product is well-defined. As an example, given an image of a face (see Figure 6.1) the local feature extraction algorithm extracts a number of image patches as the data representation. Given another image of the same face (see Figure 6.1a) the extraction is again performed - it is not guaranteed, however, that features are extracted in the exact same order because of variations in feature saliency or different backgrounds, etc. A scalar product such as in equation 6.1 would then multiply the two feature vectors as illustrated in Figure 6.1a and thus compare the "wrong" patches. In order to perform the scalar product operation in a meaningful manner one would have to first *reorder* both feature vectors such that corresponding features occupy corresponding vector indices (see Figure 6.1b).

Figure 6.1: Illustration of the correspondence problem. a) standard scalar product chooses wrong correspondences b) reordering the feature vector solves the correspondence problem



It is interesting to note in this respect that the same problem applies when using data representations consisting of raw pixel images. The effect of the correspondence problem is, however, reduced to a large degree due to a number of reasons

- usually the number of pixels (for example, 32x32 pixels is a common choice for SVM experiments) is higher than the number of local features (typically of the order of 100)
- feature extraction algorithms introduce additional measures of saliency
- neighborhood relations are preserved in an image vector

Returning to local features, it seems clear that in order to construct a well-defined local feature SVM framework both feature vector length and correspondence issues have to be addressed. However, the issue goes beyond being able to perform scalar products: in order to benefit from the advantages of large margin classifiers when using local features, it is crucial to develop novel strategies for measuring *local similarities* with scalar products in general. In other words, one needs to define a new class of kernels.

6.3 Local kernels

In this section, I want to define a new class of kernels for local features and show how these kernels can be applied to some existing approaches for matching and recognition using local features.

Definition: Denote by $K(x, y)$ a kernel function and by $\mathcal{I} = \{\vec{I}_i\}_{i=1}^m$ a set of images and $F = \{F_i\}_{i=1}^m$ the corresponding set of local features, with $F_i = \{\vec{f}_j(\vec{I}_i)\}_{j=1}^{n_i}$, $i = 1, \dots, m$. For all $(F_h, F_k) \in \mathcal{F}$, consider the function

$$K_f(F_h, F_k) = \frac{1}{2} \left[\hat{K}(F_h, F_k) + \hat{K}(F_k, F_h) \right] \quad (6.2)$$

with

$$\hat{K}(F_h, F_k) = \frac{1}{n_h} \sum_{j_h=1}^{n_h} \max_{j_k=1, \dots, n_k} \left\{ K_l(\vec{f}_{j_h}(\vec{I}_h), \vec{f}_{j_k}(\vec{I}_k)) \right\}.$$

and $K_l(\vec{f}_{j_h}, \vec{f}_{j_k})$ being any Mercer kernel.

If the local features contain position information, equation 6.2 can be extended as follows:

Definition: Denote by $\mathcal{I} = \{\vec{I}_i\}_{i=1}^m$ a set of images and $F = \{F_i\}_{i=1}^m$ the corresponding set of local features, with $F_i = \{\vec{p}_j(\mathbf{I}_i), f_j(\mathbf{I}_i)\}_{j=1}^{n_i}$, $i = 1, \dots, m$. For all $(F_h, F_k) \in \mathcal{F}$, consider the function

$$K_{fp}(F_h, F_k) = \frac{1}{2} \left[\hat{K}(F_h, F_k) + \hat{K}(F_k, F_h) \right] \quad (6.3)$$

with

$$\begin{aligned} \hat{K}(F_h, F_k) = & \frac{1}{n_h} \sum_{j_h=1}^{n_h} \max_{j_k=1, \dots, n_k} \{K_l(\vec{f}_{j_h}(\vec{I}_h), \vec{f}_{j_k}(\vec{I}_k)) \\ & \cdot \exp\{-(\vec{p}_{j_h}(\vec{I}_h) - \vec{p}_{j_k}(\vec{I}_k))^2 / 2\sigma^2\}\}. \end{aligned}$$

What equations 6.2, 6.3 effectively do is to establish *correspondences* via the sum over maximum values in the kernel function such that

- the scalar product in $K(\mathbf{x}, \mathbf{y}) = K(\mathbf{x} \cdot \mathbf{y})$ runs over corresponding features
- different feature vector lengths in F are taken into account by symmetrizing in equation 6.2.

There are two important points related to this particular design of local kernels. First, it should be noted that the \max -function was chosen for determining correspondences as it represents one of the most simple ways of matching between two local features. Whereas one might certainly introduce more sophisticated matching schemes at this point (such as voting, etc.), the proposed method is best suited for highly discriminant and robust feature descriptors (such as, for example, developed by Lowe [2004], Schmid and Mohr [1997], Mikolajczyk and Schmid [2005]). For these descriptors it can be expected that the scalar product between two corresponding features has a much lower value (ideally zero) than between two arbitrarily chosen non-corresponding features. Second, summing over all available features for both images creates a symmetric similarity metric such that $K_{f,fp}(F_h, F_k) = K_{f,fp}(F_k, F_h)$. This property thus fulfills by definition one of the requirements for a Mercer kernel by making the kernel matrix K symmetric.

In the following, I want to give a few examples of possible kernel functions, which make use of the proposed framework and of which some will be used in the computational experiments:

Example 1 Jet features [Schmid and Mohr, 1997] are a particularly successful example of local features in the literature. Similarity between jet features, which are differential intensity invariants computed around interest points, is measured via the Mahalanobis distance:

$$d_M(\vec{f}_i, \vec{f}_j) = \sqrt{\langle \vec{f}_i - \vec{f}_j | \Lambda^{-1} | \vec{f}_i - \vec{f}_j \rangle}$$

where Λ is the covariance matrix of the components. d_M can be easily mapped into an Euclidean distance d_E : the covariance matrix is a real symmetric positive semi-definite matrix, which can be decomposed via SVD:

$$\Lambda^{-1} = P^T D P,$$

with P orthogonal and D diagonal. It follows that

$$d_M(\vec{f}_i, \vec{f}_j) = d_E(\sqrt{D} P \vec{f}_i, \sqrt{D} P \vec{f}_j).$$

Thus we can use any of the following kernels as K_l in equation (6.2):

$$K_{J(p)}(\vec{f}_i, \vec{f}_j) = \left((\sqrt{D} P \vec{f}_i \cdot \sqrt{D} P \vec{f}_j) + c \right)^p, p \in \mathcal{N}, c \in \mathbb{R}^+$$

Figure 6.2: Exemplars from the COIL (top row), COGVIS-ETH (middle row) and FACE (bottom row) databases. Note the different degrees of homogeneity of the object classes.



Example 2 Schaffalitzky and Zissermann [2001] proposed to compute local histograms at different scales around detected points of interest; they compare the local features via χ^2 similarity measures. For these local features one can use as K_l the intersection measure introduced by Swain and Ballard [1991], which was proven to be a Mercer kernel [Barla et al., 2002], or

$$K_{\chi^2}(\vec{f}_i, \vec{f}_j) = \exp \{-\rho \chi^2(\vec{f}_i, \vec{f}_j)\}, \quad (6.4)$$

$$K_{a,b}(\vec{f}_i, \vec{f}_j) = \exp \{-\rho \|f_i^a - f_j^a\|^b\}, \quad (6.5)$$

with $a \in \mathbb{R}^+, b \in]0, 2]$. Both are Mercer kernels [Belongie et al., 2001, Vapnik, 1998] and have been successfully used with histogram features [Chapelle et al., 1999, Caputo and Dorko, 2003].

Example 3 In Bühlhoff et al. [2002] a first application of local SVM kernels was given, which were used on tracked local features. The kernel used was similar to \hat{K} in equation (6.2), with K_l given by

$$K_l = \exp \left\{ -\rho \left(1 - \frac{\langle \vec{f}_i - \mu_{\vec{f}_i} | \vec{f}_j - \mu_{\vec{f}_j} \rangle}{\|\vec{f}_i - \mu_{\vec{f}_i}\| \|\vec{f}_j - \mu_{\vec{f}_j}\|} \right) \right\} \quad (6.6)$$

which satisfies Mercer condition.

6.4 Experiments

This section presents recognition experiments showing that SVM and local features, combined in the local kernel framework, outperform several state-of-the-art recognition techniques used in the computer vision literature. A total of three different databases was used for recognition experiments with four different feature types (two global and two local). Although the chosen databases vary in homogeneity and types of object classes used, they all contain real-world 3D objects rotating in depth, which makes it possible to study the degree of *view generalization* of the various recognition methods. This task is especially well-suited to examine the performance of the classifiers in real-world conditions, as viewpoint rotations introduce non-trivial changes in the image. For each experiment, SVMs were benchmarked against a nearest neighbor classifier (NNC). For

each feature type, an appropriate similarity measure was chosen for both classifiers with the aim of enabling a fair comparison between all conditions.

In the following the experimental settings are described in detail (section 6.4.1). Section 6.4.2 describes and discusses recognition results on the three databases, using all feature types for both classification schemes. In order to assess the robustness of the recognition methods, Section 6.4.3 presents recognition results in the presence of noise and occlusion. These experiments were run on one database, using all feature types and again both classifiers.

6.4.1 Experimental Setup

6.4.1.1 Datasets

The COIL dataset (Murase and Nayar [1995], Figure 6.2, top row) is one of the best known benchmarks for object recognition algorithms. It consists of 7200 color images of 100 objects (72 views for object); each image is 128×128 pixels. The images were obtained by placing the objects on a turntable and taking a view every 5° . The chosen training set consisted of a subset of 17 views per object, resulting in a view every 20° .

The COGVIS-ETH dataset (Leibe and Schiele [2003], Figure 6.2, middle row) is a recently released database, consisting of 80 objects from 8 different object categories (apple, tomato, pear, toy-cows, toy-horses, toy-dogs, toy-cars and cups; each category contains 10 exemplars). Each object is represented by 41 images from viewpoints spaced equally over the upper viewing hemisphere, at distances of $22.5 - 26^\circ$. Objects are shown on a blue background without rescaling. For training a subset of the available views on the horizontal equator was used, resulting in 16 views per object, spaced at 22.5° .

The FACE dataset (Blanz and Vetter [1999], Troje and Bühlhoff [1996], Figure 6.2, bottom row) contains 100 faces (50 male, 50 female). Each image is a high-quality rendering of a laser-scanned face without hair. Face images are resized and color-equalized in order to avoid scanning artifacts. The dataset consists, for each face, of 13 views spaced 15° from left to right profile view; faces are rendered on a black background.

The test set for each database was chosen so that its views were *in between* training views. In addition, the number of training views was also varied which made it possible to investigate view generalization across larger viewing angles. Nevertheless, even for the largest training set, classifiers had to recognize objects under a depth rotation of 15° for the FACE, 20° for the Columbia and 22.5° for the COGVIS dataset, respectively. Given the size and complexity of the databases, this already represents a hard recognition problem for any classification scheme.

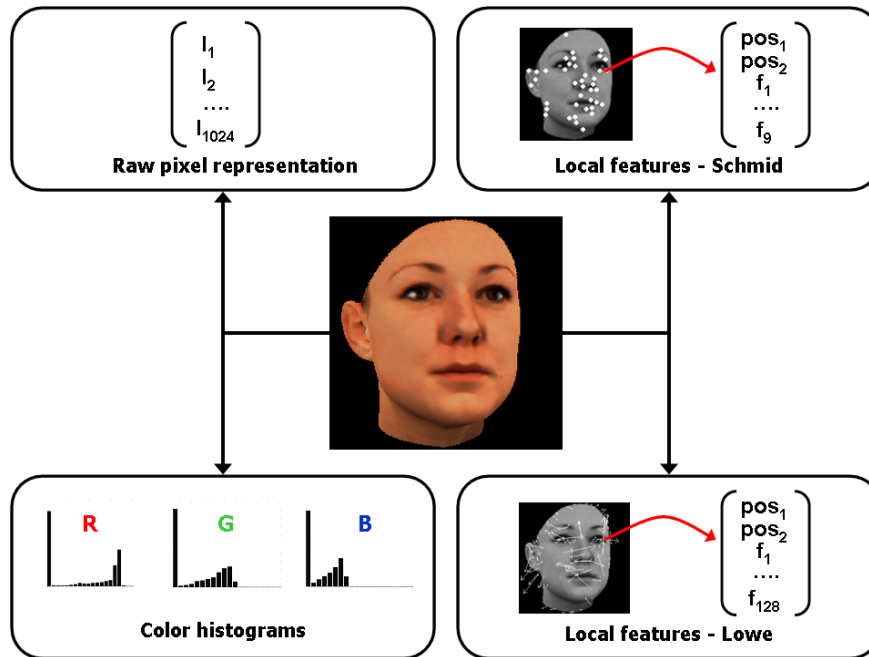
6.4.1.2 Image representations

For the experiments two global representations were used (raw pixels and color histograms) and two local representations (low-dimensional differential invariants by Schmid and Mohr [1997] and high-dimensional scale invariant feature descriptors by Lowe [2000]) (Figure 6.3).

Global representations: Raw pixel representations were extracted from the databases by conversion of all images to 32×32 pixel gray-level images, thus resulting in a vector with 1024 dimensions. Color histograms were evaluated on the full-size color images with 20 bins in each of the three color channels (R,G and B) and normalized to the bin with the highest pixel counts.

Local representations: As the first local representation, jet features proposed in Schmid and Mohr [1997] were selected, which consist of a 9-dimensional vector of differential invariants computed around a number of interest points over several scales. Detection of interest points was done using a standard Harris-type corner detector, which was shown to have high repeatability and robust performance [Schmid et al., 2000]. On average, such a representation contained around 100 features per image. The second local representation consisted of the scale-invariant feature descriptors (SIFT-keys) first introduced by Lowe [2000]. For all images the default parameters of

Figure 6.3: The four representations used in the computational experiments.



the Unix-software (available from <http://www.cs.ubc.ca/~lowe/keypoints>) were used in order to extract feature keys. Besides position and scale information, each feature key contains a 128-dimensional descriptor, which is derived from an orientation histogram of image derivatives in a local neighborhood. The average number of features for the databases was 240.

6.4.1.3 Classifiers

All SVM experiments were ran using the SVMlight software [Joachims, 2002], to which implementations of the local kernels were added. In all experiments, ρ was selected via cross-validation on the test set, with C set to the default value of 100. SVMs were benchmarked against a standard nearest neighbor classification (NNC) scheme.

Classification protocols: Since the experiments require a multi-class protocol for classification, we implemented a one-versus-the-rest scheme for training and a winner-takes-all strategy for testing (see, for example, Vapnik [2000]). It is of course possible to use more sophisticated methods for classification than NNC, such as voting schemes and decision trees [Duda et al., 2001, Schmid and Mohr, 1997, Schiele and Crowley, 2000]. However, the main rationale behind using NNC is that the multi-class SVM protocol is based on a simple winner-takes-all strategy - which is exactly a NNC. Although one could incorporate such classification methods also into the SVM classification protocol, the main focus here is rather to demonstrate how the proposed local kernels can be used "out of the box" to achieve better recognition performance.

Distance metrics and kernel functions: Distance metrics for NNC, and accordingly kernel functions for SVMs, varied with representations. For the raw pixel representation, a standard Euclidean distance was used, that is

$$d_E(\vec{f}_i, \vec{f}_j) = \sqrt{\langle \vec{f}_i - \vec{f}_j | \vec{f}_i - \vec{f}_j \rangle}$$

The corresponding kernel (in the sense that it is the kernel which maps the data into an Euclidean space [Burges, 1998]) is the Gaussian kernel (equation 6.5), with $a = 1, b = 2$.

For color histograms the standard χ^2 distance was chosen:

$$d_{\chi^2}(\vec{f}_i, \vec{f}_j) = \sum_i \frac{(\vec{f}_i - \vec{f}_j)^2}{\vec{f}_i + \vec{f}_j}$$

where the corresponding kernel is the Gaussian kernel (equation 6.4). Finally, the following distance metric was used for comparing local features:

$$d_L(\vec{f}_i, \vec{f}_j) = \sum_i \max_{j=1, \dots, n_k} \frac{\langle \vec{f}_i - \mu_{\vec{f}_i} | \vec{f}_j - \mu_{\vec{f}_j} \rangle}{\|\vec{f}_i - \mu_{\vec{f}_i}\| \|\vec{f}_j - \mu_{\vec{f}_j}\|}.$$

Note that this metric does not make use of information contained in the feature positions, such as local feature constellations [Schmid and Mohr, 1997], or global feature layout [Bülthoff et al., 2002]. This experimental setup was chosen in order to examine the usefulness of the local features themselves. The corresponding kernel is given by (6.6), with $K_l = \frac{\langle \vec{f}_i - \mu_{\vec{f}_i} | \vec{f}_j - \mu_{\vec{f}_j} \rangle}{\|\vec{f}_i - \mu_{\vec{f}_i}\| \|\vec{f}_j - \mu_{\vec{f}_j}\|}$.

6.4.2 Experimental Results: View Generalization

Tables 6.1 report error rates (e.r.) for the three databases as a function of depth rotation generalization, feature type and classifier. From the analysis of these results, we can draw the following main conclusions:

- Regardless of the data representation, SVMs show large performance improvements compared to NNC. As the experimental setup is identical for both classifiers, this provides further evidence for the superior generalization properties of SVMs.
- This performance improvement is highest for small depth rotations and lowest for generalization across large depth rotations with the most extreme example given by results on generalization across 15° for the face database.
- SVMs with the proposed local kernel achieve the best performance on all databases for depth rotations of up to 45° . This result shows the benefit of introducing highly discriminant local feature representations within the SVM framework for object recognition tasks.
- As could be expected, classification errors generally increase with increasing amount of depth rotation. This increase is much more pronounced for the face database showing how difficult face recognition remains despite the relatively good results for the "object" databases.
- For depth rotations up to 45° local representations outperform global representations demonstrating the advantage of local analysis for smaller changes in pose.
- For larger view rotations, color histograms yield significantly less recognition errors than both raw pixel and local representations. This agrees with earlier results by Swain and Ballard [1991], but is nevertheless surprising due to the low dimensionality of the chosen histograms. Taken together, these two results seem to suggest that a hybrid strategy for object recognition might be successful - where correspondences for local features are hard to find due to large changes in pose, global measures can still give significant information about the object.
- For local representations, SIFT yields higher recognition performance than Jet Features, which can be attributed to their increased discriminability at the price of a much higher number of dimensions (128 dimensions compared to 9 dimensions for Jet Features).

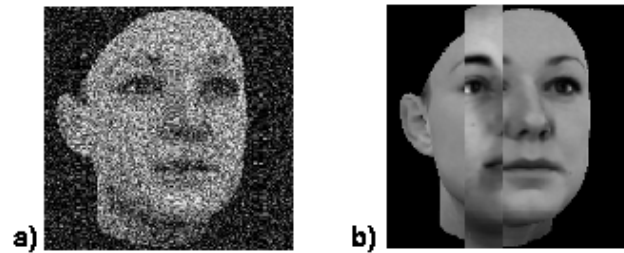
COIL				
Generalization	20°	40°	60°	80°
Raw pixel representation				
NNC	16.7	40.6	30.0	61.3
SVM	5.9	26.3	16.3	57.5
Color histogram representation				
NNC	4.8	18.4	11.6	41.6
SVM	3.5	12.0	9.0	39.5
Jet Feature representation				
NNC	11.7	46.2	42.0	68.0
SVM	1.5	29.2	29.7	64.2
SIFT representation				
NNC	6.4	27.0	36.0	53.0
SVM	3.1	14.3	16.7	50.0

COGVIS				
Generalization	22°	45°	68°	90°
Raw pixel representation				
NNC	18.9	33.8	34.2	63.2
SVM	10.2	13.7	16.3	58.2
Color histogram representation				
NNC	21.3	27.2	32.3	65.0
SVM	5.9	19.7	22.4	61.9
Jet Feature representation				
NNC	11.4	36.3	67.5	75.0
SVM	1.6	22.7	45.3	71.0
SIFT representation				
NNC	11.3	47.8	50.8	78.8
SVM	2.1	19.1	30.5	76.3

FACE						
Generalization	15°	30°	45°	60°	75°	90°
Raw pixel representation						
NNC	65.0	89.7	98.0	97.0	95.0	97.0
SVM	3.2	82.5	92.5	96.0	96.0	95.5
Color histogram representation						
NNC	6.8	39.3	50.3	80.0	86.0	90.0
SVM	2.7	33.0	38.0	77.5	85.0	89.0
Jet Feature representation						
NNC	1.2	34.3	84.0	95.0	97.0	99.0
SVM	0.0	28.6	79.3	94.7	95.0	99.0
SIFT representation						
NNC	0.0	17.0	57.0	88.0	99.0	99.0
SVM	0.0	6.0	18.0	86.0	93.5	98.0
SIFT representation + SIFT matching						
NNC	0.0	12.5	39.4	87.0	98.0	100.0

Table 6.1: Classification errors in percent on three databases for global and local representations, using SVM and NNC.

Figure 6.4: Two types of image degradation: a) Gaussian Noise and b) Occlusion



6.4.3 Experimental Results: SIFT versus Local Kernels

In a second series of experiments, the influence of the matching function defined in equation 6.2 was investigated. To this end, the SIFT representation was also tested with the matching algorithm delivered with the software package. For matching two features, this algorithm uses an additional constraint on the goodness of the *second-best match* in order to decrease the number of false positives: if the best match is larger than 0.6 times the distance to the second closest it is rejected. This constraint usually results in a smaller number of very reliable and discriminative feature matches, as each correspondence relies on two good matches.

The bottom row of Table 6.1 lists the results on the face database for this matching algorithm. For viewing angles up to 45° , the additional constraint yields better recognition rates than the simple implementation of a greedy \max -function combined with NNC. However, this version of the algorithm is still significantly outperformed by the combination of SVMs with the proposed local kernel.

6.4.4 Experimental Results: Recognition under Noise

By adding two types of image degradations (noise and occlusion) to the *test set* of the FACE database the robustness and generalization capabilities of the proposed local kernel framework was tested. The FACE database was selected for these experiments because here SVM and NNC yield very similar performance on local features (see Table 6.1).

First of all, Gaussian noise of 10% strength was added to each image² (Figure 6.4a), which is a manipulation of the global statistics of the image. The second type of image degradation, consisted of masking out a random part of the image by inserting data from a different image (Figure 6.4b), which represents a more local disruption of image statistics. The portion of the face that was occluded was set to 15% of the image size. The task for the classifier is thus to recognize objects *both* under depth rotations and additional noise or occlusion.

The results shown in Table 6.2 demonstrate that recognition performance has decreased for all representations and classifiers; but once again SVM performed better than NNC, for both kinds of degradation and all feature types. With respect to local features, these results thus confirm their improved robustness to noise and occlusion (Table 6.2, lower two blocks).

Raw pixels: results of both classifiers reasserts that this representation is not robust in presence of noise or occlusion.

Color Histograms: adding Gaussian Noise severely disrupts color information in all three channels, which leads to an extremely poor performance for both classifiers (Table 6.2, left column, middle). As already observed in Swain and Ballard [1991], color histograms seem to be relatively more robust to occlusion (Table 6.2, right column, middle). However, performance still drops about

²Note that this applies to all three color channels.

Algorithm	Noise	Occlusion
Raw pixel representation		
NNC	76.7	89.1
SVM	35.0	58.2
Color histogram representation		
NNC	98.7	68.6
SVM	98.5	53.2
Jet Feature representation		
NNC	22.4	38.2
SVM	5.4	26.7
Jet Feature representation & position constraint		
NNC	9.5	34.2
SVM	1.4	13.2

Table 6.2: Classification errors in percent on the FACE dataset in presence of noise or occlusion.

50-60% compared to the uncluttered condition (see Table 6.1, right column, middle).

Local Features: Local features performs quite well under noise, but suffer from occlusion (although much less than global features). Again, SVMs with the local kernel significantly outperforms NNC in both conditions.

6.4.5 Experimental Results: Recognition using position constraints

Finally, recognition results on a different kernel (and corresponding metric), which includes a global position constraint in the form of eq. 6.3 (see also Bühlhoff et al. [2002]), show that it is possible to significantly improve recognition performance on degraded images by incorporating this extra information. Results for the local representation on the face database (Table 6.2, last two rows) show that, by using a position constraint, improvements of up to 13% for both NNC and SVM are possible. Recognition performance is still much better under noise than under occlusion; one of the reasons for this is that a fairly global position constraint was used in this case (see also chapter 4). Other types of more local position constraints (as done in Schmid and Mohr [1997] for instance) might be more appropriate in this case.

6.5 Conclusion

In this section, I proposed a "recipe" for constructing kernels which are suitable for object recognition with local features. Several examples of local kernels suitable for different types of local features discussed in the literature were given, which demonstrates that these types of kernels can be useful for a wide range of applications in the computer vision community. In addition, experiments on three different databases were presented which compared global versus local features, using NNC and SVM. In all cases, SVM resulted in a significant increase in performance, which again confirms the advantage of large-margin classifiers regardless of the underlying data representation. Moreover, recognition results obtained using local features combined with SVM outperform both NNC with local features as well as SVMs with global representations. In the following chapter, I will investigate the advantages of explicitly using spatio-temporal representations in a SVM framework.

Chapter 7

Computational studies III - SVMs and keyframes

Traditionally object recognition - at least in the area of computer vision - has relied mostly on static representations (see, for example, Pontil and Verri [1998], Roth et al. [2002], Blanz et al. [2002] for SVM-related recognition studies). One of the central topics of this thesis has been that natural visual input consists of *spatio-temporal* patterns - illustrated by the results from psychophysical studies, which corroborate that the human visual system is able to exploit inherent temporal characteristics for recognition processes [Wallis and Bühlhoff, 1999, 2001, Wallis, 2002]. In this chapter, the focus lies on combining *spatio-temporal* data representations from the area of computer vision with robust recognition schemes from machine learning. In particular, I will present a framework on how to combine the keyframe framework with support vector machines as an extension of the previous chapter, where only static information was used.

From a computer vision perspective, the most simple data representation for analysis of an image *sequence* consists of the raw pixels of each individual frame. Since taking every frame of the sequence, however, would be computationally intractable for most practical applications, a straightforward strategy could reduce input data size by taking only every *n*th frame. Choosing the proposed keyframe framework instead of such a more inflexible representation will thus result in a much more efficient and robust representation for sequence analysis as I have shown in chapters 4 and 5.

In the following I want to briefly recapitulate the main concepts of the keyframe framework as it is used here (see also Figure 7.1): as a first step, local visual features are extracted in each image, which consist of small image patches centered around so-called interest points (in our case, corners). Each visual feature thus carries position information (the pixel position in the image) as well as appearance-based information (the image patch). In order to process image sequences, this initial set of visual features is now tracked over the subsequent images of the input sequence. Since features inevitably will get lost during tracking due to changes in the visual content of the depicted object, the process is restarted at these points in time. The final representation then consists of the set of frames, at which tracking had to be restarted and thus implicitly incorporates the *a priori* knowledge of a sequential image presentation (temporal continuity, see chapter 1 and Wallis and Bühlhoff [1999, 2001], Wallis [2002]). Alternatively one could see this as an automatic segmentation process, which divides the input sequence into temporally coherent chunks of which only the endpoints are kept (see chapter 3). In addition, the tracking process itself provides further temporal information through the *feature trajectories* which encode the feature transformation between two consecutive keyframes. Thus, feature tracking and sequence segmentation allows for construction of several spatio-temporal data representations capturing different degrees of spatio-temporal information. One main focus of this chapter is to assess the generalization performance of these various data representations ranging from raw image pixels to full spatio-temporal feature representations in a demanding object recognition task.

In a recent study by Stringer and Rolls [2002], learning of feature representations in a temporal context was achieved by a hierarchical network model. The model provided a neural network implementation for learning of such transformations and was tested on simple artificial images. Motivated more from the computer vision perspective, our approach is to derive these feature transformations *explicitly* from the input data and in addition to test a variety of spatio-temporal representations on realistic high-dimensional input data.

The second focus of this chapter lies on robust classification schemes from statistical learning theory, in particular on Support Vector Machines (SVMs, Vapnik [2000]) in a normalized feature space (see Graf et al. [2003]). Of the several studies which have used SVMs in an object classification task, some of the most relevant in our context are: Pontil and Verri [1998], who conducted the first extensive study on image classification using SVMs, Chapelle et al. [1999], where classification was done using novel similarity metrics based on image histograms, and Heisele et al. [2001], who developed a framework for part-based classification and detection of faces using visual features. Here we extend this line of research on classification of images by developing metrics suitable for processing the spatio-temporal feature representations outlined above with SVMs. Again, the main reason for the need for novel metrics is that - similarly to the previous chapter - the structure of these data representations does not allow straightforward integration with kernel functions—in particular, standard scalar products for comparing two input vectors cannot be applied in this case. Indeed, the proposed similarity metrics can be integrated not only into SVMs but into all types of kernel machines.

The remainder of the chapter is structured as follows: section 7.1 presents an overview of the algorithms on feature extraction, tracking and sequence segmentation, derives several spatio-temporal data representations from this and finally details the similarity metrics and kernel functions necessary for processing these representations. Section 7.2 describes the database of image sequences and presents computational recognition experiments comparing the various spatio-temporal data representations both for multi-class SVMs and nearest-neighbor classifiers.

7.1 Algorithmic Overview

In this section I want to briefly recapitulate the keyframe representation and cast it into a context suitable for integration into a support vector framework. Again, the crucial point is to develop suitable similarity metrics taking into account correspondences between local feature sets.

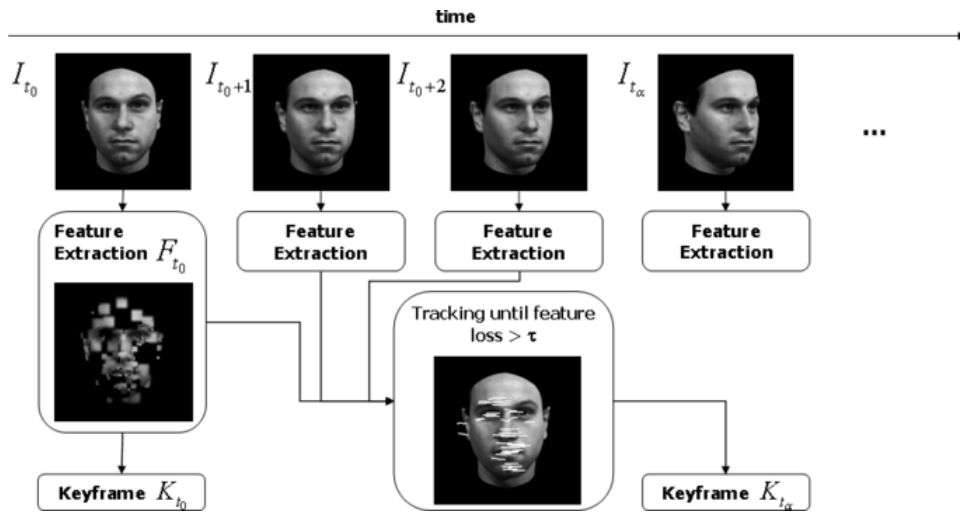
7.1.1 Image Sequence Processing

Evidence from psychophysical studies shows that humans are able to exploit the temporal continuity in the natural visual input during learning of object classes [Wallis and Bühlhoff, 2001, Wallis, 2002]. A computational framework directly motivated by these results is introduced in Wallraven and Bühlhoff [2001a], where a sparse yet powerful representation is generated from spatio-temporal visual input. The main idea behind this approach, which we extend here, is to process a sequence of images $\mathcal{I} = \{I_t\}_{t=t_0}^{t_1}$, $t \in \mathbb{N}$ to extract salient features and track them in subsequent frames in order to segment the sequence.

7.1.1.1 Feature Extraction

As a first step, extraction of salient point features is done at multiple image scales by a standard interest point detector (see chapter 3). A set of *appearance-based features* $F_t = \{\vec{p}_i(I_t), \vec{f}_i(I_t)\}_{i=1}^n$ is then constructed by using the pixel positions $\vec{p}_i(I_t)$ of the interest points together with their pixel neighborhood \mathcal{N} written in vector form as $\vec{f}_i(I_t)$. The extension to other types of feature descriptors can of course be done in the same manner as in the previous chapter. The focus here, however,

Figure 7.1: Overview of the keyframe system as used in this chapter (cf. Figure 3.1).



is not on the feature descriptor itself but rather on the spatio-temporal information contained in the final representation of the image sequence.

As shown in Figure 7.1, extracted visual features tend to cluster at facial features such as the eyes, mouth, nose, etc. This clustering is due to the fact that these features usually exhibit much higher variation in pixel intensities than, for example, the cheeks or the forehead. Thus, feature extraction seems to largely focus visual processing on regions in the image, which correspond to “real” object features (see also chapter 4).

7.1.1.2 Sequence Segmentation through Tracking

Sequences are processed by tracking the initial set of features at time t_1 across subsequent images until the percentage of features that could be continuously tracked drops below a certain threshold τ (see Figure 7.1). At this point in time (t_α), a new “keyframe” (K_{t_α}) is inserted into the representation and the process is restarted until the sequence ends. This process basically partitions the sequence into segments in which a temporally continuous feature motion occurred.

The final representation of the sequence thus consists of a set of keyframes \mathcal{K} , each of which contains a set of salient visual features as described previously. One way of explicitly capturing the *spatio-temporal* information that has led to the formation of these keyframes is to analyze the trajectories of tracked features between keyframes. As shown in Figure 7.1, for a face executing a horizontal rotation, the extracted feature trajectories directly encode this horizontal object motion between keyframes. Thus, a straightforward way of adding this information to the representation is to determine an image-centered *direction vector* $\vec{v}_{\alpha\beta}$ of unit length, which encodes the average feature motion between keyframes K_{t_α} and K_{t_β} :

$$\vec{v}_{\alpha\beta} = \frac{\sum_{i=1}^l [\vec{p}_i(K_{t_\alpha}) - \vec{p}_i(K_{t_\beta})]}{\left\| \sum_{i=1}^l [\vec{p}_i(K_{t_\alpha}) - \vec{p}_i(K_{t_\beta})] \right\|}$$

where i indexes all of the l feature trajectories that could be tracked from K_{t_α} to K_{t_β} . From this segmentation process it is now possible to derive four different representations with which the input image sequence can be described and which carry different amounts of spatio-temporal information:

- An image representation consisting of the raw pixel data of the keyframes \mathcal{K}_I . This is the most basic view-based representation of the sequence.

- A feature representation consisting of only the appearance information of the features \mathcal{K}_f . This representation encodes the visual information contained in each keyframe by a number of salient interest points and thus concentrates visual information at object features (see Figure 7.1).
- A feature representation consisting of appearance information and additional spatial information in the form of the image positions of the features \mathcal{K}_{fp} . This basically provides access to the 2D spatial layout of features in the image.
- The full spatio-temporal representation \mathcal{K}_{fpv} , which consists of appearance and spatial information as well as the temporal information from the direction vectors, which was derived from the feature trajectories. In addition to encoding visual interest points, this representation makes the *temporal* formation of the image sequence most explicit.

Below, a general framework for processing of the feature representations as well as appropriate similarity metrics for their integration with SVMs is presented.

7.1.2 Feature Matching

In this section, the general framework for matching two sets of visual features is outlined. In addition to the tracking process, which forms the basis for sequence segmentation and keyframe generation, this framework will also be applied to matching the keyframes themselves, that is, for recognition of novel sequences or images in general.

The algorithm (see also Wallraven and Bülthoff [2001a], Scott and Longuet-Higgins [1991], Pilu [1997]) matches two feature sets of two images, F_{t_1} and F_{t_2} , by constructing a pairwise similarity matrix \mathbf{A} given by:

$$A_{ij} = \delta_f(\vec{f}_i(I_{t_1}), \vec{f}_j(I_{t_2})) \cdot \exp(-\rho \delta_p(\vec{p}_i(I_{t_1}), \vec{p}_j(I_{t_2})))$$

where the number of features in the two images $i = 1, \dots, n$ and $j = 1, \dots, m$ is not necessarily equal ($n \neq m$). Each entry A_{ij} in the similarity matrix is composed of two terms:

- An appearance similarity measure δ_f , which depends solely on \vec{f} . This can be, for example, the normalized cross-correlation $\delta_f = \frac{\langle \vec{f}_i - \vec{\mu}_i | \vec{f}_j - \vec{\mu}_j \rangle}{\sigma_i \cdot \sigma_j}$, where $\vec{\mu}$ is the mean of \vec{f} and σ the standard deviation of \vec{f} .
- A spatial similarity measure δ_p , which depends solely on \vec{p} . One of the simplest similarity measures is based on the standard Euclidean distance $\delta_p = \|\vec{p}_i - \vec{p}_j\|^2$

One of the strengths of this approach is the combination of an appearance-based term derived from image intensities *and* a spatial layout term based on positions of the features in the image plane. As will be seen in the computational experiments, this combination increases the robustness of the matching process considerably.

In addition, the parameter ρ allows to specify a *prior* on the expected spatial relationships between the two feature sets. For tracking purposes, this parameter is kept high in order to penalize large distances between two features, which has the effect of constraining matches to small changes in spatial layout. Accordingly, for general purpose matching of two feature sets, ρ is set to lower values in order to allow for a larger change in feature transformations.

The similarity matrix \mathbf{A} is used to find a one-to-one feature mapping between two feature sets F_{t_1} and F_{t_2} . In a first step, it finds the largest element in \mathbf{A} , that is, the feature pair with the highest similarity value. The row and column defined by this pair are then deleted from the matrix and the process is repeated until a pre-determined number Q ($0 < Q \leq \min(n, m)$) of corresponding feature pairs have been found. By *summing up the similarity values* of the corresponding feature pairs, a matching score can be determined, which represents the quality of the match $\mathcal{M}(\mathbf{A})$ between F_{t_1} and F_{t_2} .

7.1.3 Image and Feature Matching for Kernel Machines

In this section, we describe similarity metrics M , which have to be incorporated into the kernel functions in order to process either the image or the spatio-temporal feature representations. Whereas image similarity metrics can be based on straightforward Euclidean distances, kernel machines require novel feature similarity metrics.

7.1.3.1 Similarity Metrics for the Image Representation

In line with the experiments conducted by Pontil and Verri [1998], comparison of two images I_t and I_u (the use of a different time index u indicates that the second image might come from another image sequence) is based on the pixel-based Euclidean norm, which is evaluated directly on the vector of image intensities:

$$M_I(\vec{I}_t, \vec{I}_u) = \|\vec{I}_t - \vec{I}_u\|^2$$

Note that this metric is highly sensitive to all kinds of changes in image intensities introduced by illumination, rotation in depth, etc.

7.1.3.2 Similarity Metrics for the Feature Representation

The proposed feature representation, as introduced in the previous section, cannot be integrated with kernel machine classifiers directly. To illustrate this, let us assume two sequences, $\mathcal{I} = \{I_t\}_{t=t_0}^{t_1}$ and $\mathcal{J} = \{J_u\}_{u=u_0}^{u_1}$. Disregarding the direction vector for the moment, the corresponding feature representations, that is, position and feature vectors of the *keyframes*, as defined above are:

$$\mathcal{K}_{fp}(\mathcal{I}) = \{F_{t_\alpha}\}_{t_\alpha} = \{\vec{p}_{n_{t_\alpha}}(K_{t_\alpha}), \vec{f}_{n_{t_\alpha}}(K_{t_\alpha})\}_{t_\alpha}$$

and

$$\mathcal{K}_{fp}(\mathcal{J}) = \{F_{u_\alpha}\}_{u_\alpha} = \{\vec{p}_{m_{u_\alpha}}(K_{u_\alpha}), \vec{f}_{m_{u_\alpha}}(K_{u_\alpha})\}_{u_\alpha}$$

where α indexes all keyframes in the two sequences and $n_{t_\alpha}, m_{u_\alpha}$ is the number of local features in each keyframe.

In general, the task is to match a feature set F_{t_α} from $\mathcal{K}(\mathcal{I})$ with a feature set F_{u_α} from $\mathcal{K}(\mathcal{J})$. However, as the features originate from a particular input sequence with a particular set of tracked features, this poses two major problems. First, the number of features ($n_{t_\alpha}, m_{u_\alpha}$) for each keyframe is not the same, which makes straightforward scalar products of the form $\langle F_{t_\alpha} | F_{u_\alpha} \rangle$ —and with this also the kernel functions based upon these— between two feature sets ill-defined. However, even if the number of features were equal (for example, by constraining it manually), it cannot be guaranteed that the *order* of the features within each set is the same in both representations. Taking scalar products between two feature representations then would be problematic because of lacking *correspondence* between vectors (see chapter 6).

One way to solve these two problems is to replace the scalar product between the feature representations of F_{t_α} and F_{u_α} with the similarity score $\mathcal{M}(\mathbf{A})$ defined in section 7.1.2¹:

$$\langle F_{t_\alpha} | F_{u_\alpha} \rangle \rightarrow \mathcal{M}(\mathbf{A})$$

The similarity between two feature sets is thus assessed based on their similarity score instead of evaluating the scalar product directly on the data. As the proposed similarity score is derived by explicitly determining *correspondences* between feature sets, it is thus possible to process two

¹In this chapter, I have introduced a slightly different notation for the solution to the local feature matching problem than in chapter 6. One of the main differences is that the solution presented here introduces the parameter Q , which makes the matching functions symmetric rather than summing over both left-right and right-left matches.

feature sets containing different numbers of features and to ignore effects of different ordering of the features.

For the classification of the three feature representations introduced above, the following similarity metrics are proposed. Using only feature information the first similarity metric is defined as:

$$M_f(F_{t_\alpha}, F_{u_\alpha}) = \mathcal{M}(\mathbf{A}^f)$$

where each entry in the matrix \mathbf{A}^f is of the form:

$$A_{ij}^f = \delta_f(\vec{f}_i(I_t), \vec{f}_j(J_{u_\alpha})) = \frac{\langle \vec{f}_i - \vec{\mu}_i | \vec{f}_j - \vec{\mu}_j \rangle}{\sigma_i \cdot \sigma_j}$$

Using both feature *and* position information the metric can be extended to:

$$M_{fp}(F_{t_\alpha}, F_{u_\alpha}) = \mathcal{M}(\mathbf{A}^{fp})$$

$$A_{ij}^{fp} = (\delta_f(\vec{f}_i(I_{t_\alpha}), \vec{f}_j(J_{u_\alpha})) + 1) \cdot \exp(-\rho \delta_p(\vec{p}_i(I_{t_\alpha}), \vec{p}_j(J_{u_\alpha}))) - 1$$

The additional position term (which is of the form $\delta_p = \|\vec{p}_i - \vec{p}_j\|^2$) has the effect of biasing the matching score towards closer matches and represents a straightforward way of including information about the *layout* of the local features in the image during matching.

Finally, when considering appearance and position information as well as the temporal prior on feature transformation, the following metric is proposed:

$$M_{fpv}(F_{t_\alpha}, F_{u_\alpha}) = \mathcal{M}(\mathbf{A}^{fpv})$$

$$A_{ij}^{fpv} = (\delta_f(\vec{f}_i(I_{t_\alpha}), \vec{f}_j(J_{u_\alpha})) + 1) \cdot \exp(-\rho \langle \vec{v}_{\alpha\beta} | \vec{p}_i(I_{t_\alpha}) - \vec{p}_j(J_{u_\alpha}) \rangle^2) - 1$$

Here, additional information derived from the keyframe trajectories is included by taking the inner product between the direction vector and the potential matching vector for each feature pair. This effectively results in biasing matches towards the *preferred direction* of feature transformation².

7.1.4 Multi-class SVMs

Support Vector Machines as binary classifiers have attracted much attention because their thorough mathematical foundations rooted in statistical learning theory [Vapnik, 2000, Schölkopf and Smola, 2002]. They are here considered in a normalized feature space $\frac{k(\vec{x}, \vec{y})}{\sqrt{k(\vec{x}, \vec{x})k(\vec{y}, \vec{y})}}$ and with a modification of the SV algorithm as introduced in Graf et al. [2003].

In the context of object recognition an extension of the binary recognition scheme is required, as most real-world computer vision problems contain a large number of classes. As shown in several studies, this is not a trivial problem (see for example Phillips [1999] and Hsu and Lin [2002] for a recent performance study of different multi-class frameworks for SVMs). Here, a standard approach is used, which does not modify the SV formulation (as opposed to Weston and Watkins [1999]). Assuming the training database contains L classes, L classifiers are first trained using a one-vs-rest protocol in which each class is trained against all other classes. For each element of the testing subset of the database, recognition is then done according to a *winner-takes-all* strategy based on the output of the decision function of all L classifiers [Vapnik, 2000, Schölkopf and Smola, 2002]. The recognition error in this case is then simply the mean of the individual recognition errors of each element of the testing set.

In the SVM framework, the kernel matrices need to be symmetric positive definite in order to satisfy Mercer's conditions. Since the metrics for feature recognition presented above are clearly symmetric in their arguments, the resulting kernel functions are also symmetric. In addition, section 3.4 presents an empirical study demonstrating that the resulting kernel matrices are positive definite under the experimental conditions outlined below.

²This corresponds closely to the geometric matching approach geo1 introduced in section 5.2.3.

7.2 Computational Experiments

This section presents object classification results on a large database of face sequences, which compare the various image and spatio-temporal feature representations using multi-class SVMs. The training set in our case consists of a set of keyframes (views), which is extracted from a sequence of a rotating face. The test set, which contains different views of the face is then used to investigate the classification performance in an identification task.

7.2.1 Database and Representation

The database used 100 sequences generated from the MPI face database, each of which showing a horizontal 180° rotation (from left profile view -90° to right profile view 90°) of the faces in 9° steps. Individual frames consist of 256x256 pixel, grayscale images with a black background. The experimental procedure for training and testing is as follows. First of all, sequences are processed with the keyframe framework introduced in section 2. For a given tracking threshold and a given tracking prior³, a set of keyframes is generated for each of the 100 sequences. These keyframes then define the training set.

As all sequences depict a controlled horizontal rotation, one would expect the generated keyframes to be equally distributed across angles. Indeed, for the chosen tracking parameters, each of the 100 sequences generated five keyframes. The average rotation angles of these five keyframes were: $-90^\circ, -50.3^\circ, -14.5^\circ, 15.1^\circ, 54.6^\circ$. The training set thus consisted of the following five views, which were chosen as the nearest views from the database: $-90^\circ, -54^\circ, -18^\circ, 18^\circ, 54^\circ$ (see middle box in Figure 7.2 for an example). In addition, the analysis of the feature trajectories yielded the following direction vectors:

$$v_{1,2} = \begin{pmatrix} 0.98 \\ 0.19 \end{pmatrix}, v_{2,3} = \begin{pmatrix} 0.99 \\ 0.14 \end{pmatrix}, v_{3,4} = \begin{pmatrix} 0.99 \\ 0.01 \end{pmatrix}, v_{4,5} = \begin{pmatrix} 0.98 \\ 0.19 \end{pmatrix}$$

All vectors have large x components, which indicates a strong horizontal motion component between keyframes.

The test set was chosen to study the *view generalization* capabilities of the classifiers and consisted of the four *intermediate* views between the keyframes, namely $-72^\circ, -36^\circ, 0^\circ, 36^\circ$ which yields 4 testing vectors for each sequence (see lower box in Figure 7.2). In total there are thus $5 \cdot 100$ training vectors and $4 \cdot 100$ testing vectors.

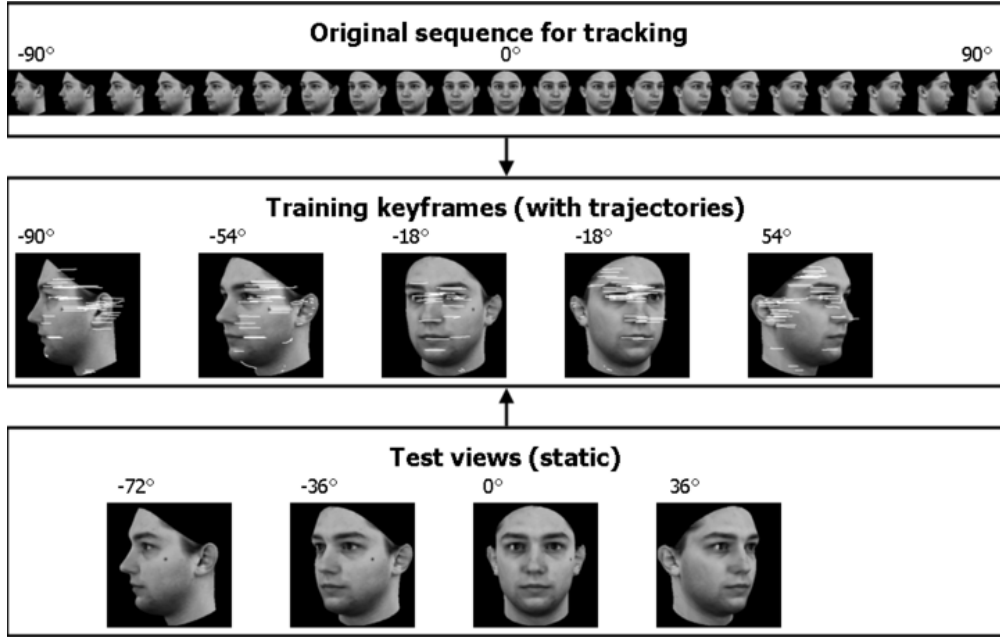
From these training and test images, we now derive four data representations. The simplest representation of the above image sequences consists of the raw pixel data from the *views* as defined above. In this case, images are reduced to $32 \cdot 32$ pixels with training and testing performed on the $32 \cdot 32$ -dimensional vector obtained from the corresponding image matrix.

Local feature representations are defined by feature tracking and keyframe extraction according to section 7.1.1, resulting in five feature sets F_{t_α} (including the direction vectors) used for training. In order to keep training and testing set separate, the four testing feature sets were extracted only on the *single static view* without taking into account any temporal information from tracking. We chose this particular testing paradigm in order to investigate the usefulness of the spatio-temporal data representations in recognizing single images⁴. For all local feature representations the size of the local pixel neighborhood \mathcal{N} around each interest point was set to 7×7 pixels and the number of features to $n = m = 20$ (this was done because of performance reasons and in no way affects the general argument made in section 7.1.1). For each keyframe this then results in a feature set F_{t_α} consisting of 1022 dimensions for the training set. This number of dimensions derives from appearance ($20 \cdot 7 \cdot 7$ dimensions), position ($20 \cdot 2$ dimensions) and direction vector

³For these tracking experiments, we used $\tau = 20\%$, $\rho = \frac{1}{20 \cdot 20}$ and $\theta = 0.2$ - see also chapter 5 for more experiments on keyframe extraction.

⁴An extension to classification of temporal sequences, however, is straightforward

Figure 7.2: Example of training and testing sets for one head.



k	1	2	3	4	5	6
recognition error	88.00%	37.50%	69.50%	81.25%	84.75%	88.50%

Table 7.1: k -NN recognition on images as a function of the number of nearest neighbors k .

information (2 dimensions). The particular parameters and number of features were chosen to make the dimensionality of both image and feature representations comparable.

7.2.2 Classification of Images

This section presents results from image classification using the raw pixel representation on a benchmark classifier, the K -nearest-neighbors (K -NN). This version of the K -NN classifier assigns an unknown pattern \vec{z} to the class $1 \leq C(\vec{z}) \leq L$ according to the minimum of the sum of the distances to the K nearest neighbors in each class: $C(\vec{z}) = \arg \min_{i=1, \dots, L} \sum_{k=1}^K \|\vec{z} - \vec{\eta}_k^i\|^2$ where $\vec{\eta}_k^i \in \mathcal{N}_E(\vec{z})$ is an element of the dataset in the Euclidean neighborhood of \vec{z} , k indexing the K nearest neighbors and i the classes.

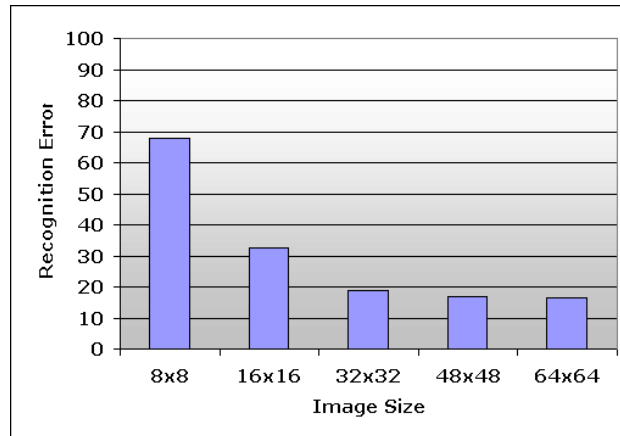
Substituting the similarity metric $M_I(\vec{I}_{t_\alpha}, \vec{J}_{u_\alpha})$, the classification error as a function of the number of nearest neighbors K is shown in Table 7.1. The optimal performance of K -NN was found at $K = 2$. This result reflects the selection of the training and testing sets. The views in the testing set are *in-between* the training views such that $K = 2$ naturally results in the best generalization performance. Note that since 100 classes are considered, chance level, that is, the mean random classification performance, is at 99% error.

The computational results reported in Table 7.2 are obtained using multi-class SVMs with RBF and polynomial kernels, where optimal SVM parameters (including the trade-off parameter C , the polynomial degree d and the RBF parameter γ) were determined using cross-validation on the training set. Not surprisingly, SVMs show a large performance increase when compared to K -NN classification, which illustrates their superior generalization capabilities for this task. Performance is roughly equal for both types of kernel functions with a slight advantage for RBF kernels.

One might argue that the severe size reduction of the images resulted in loss of information,

$k(\vec{I}_t, \vec{J}_u)$	optimal parameter	recognition error
$\exp(-\gamma \ \vec{I}_t - \vec{J}_u\ ^2)$	$\gamma = 26e - 8$	18.00
$(1 + \langle \vec{I}_t \vec{J}_u \rangle)^d$	$d = 4$	19.00

Table 7.2: SVM recognition on images for RBF and polynomial kernel functions.

Figure 7.3: Recognition error on images as a function of input image size for a polynomial kernel $(1 + \langle \vec{I}_t | \vec{J}_u \rangle)^d$.

such that the task was simply too difficult for the classifiers. Figure 7.3 shows classification results with a polynomial kernel function (with d optimally chosen), from which one can conclude that performance is bounded from below by $\approx 17\%$. Increasing the image size thus does not seem to add further discriminative information to the classification process, which suggests that the information content of the raw pixel representation is already captured well enough at $32 \cdot 32$ pixels for our database. In addition, computation time for training and testing with such high-dimensional data quickly becomes prohibitive for larger image sizes.

In summary, the results show that the chosen classifications problem seems to be a challenging task for a raw pixel representation. SVMs are able to improve the classification performance to a large degree, which provides further evidence of their strong generalization capabilities. Nevertheless, the *absolute* performance levels are still not quite satisfactory, considering that in the best test case, 18% (that is, 65 out of 400) of the test images were misclassified.

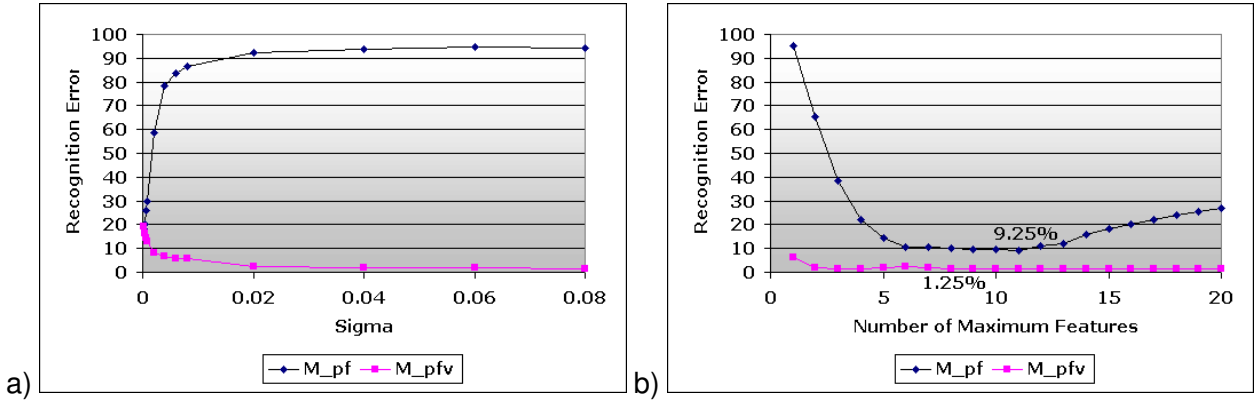
7.2.3 Classification of Features

In this section, we present classification results for the various local feature representation for benchmark K -NN and state-of-the-art SVM classifiers. As mentioned before, the nature of the interpolation problem considered here implies a minimal classification error for the K -NN classifier at $K = 2$. The benchmark experiments using K -NN classifiers are thus done for this case with the results reported in Table 7.3⁵. Comparisons between feature sets were done using the various feature similarity metrics ($M_{f,fp,fpv}(F_{t_\alpha}, F_{t_u})$) introduced in section 2.3.

The results demonstrate a dramatic increase in classification performance with respect to the raw image representation, which in this case even surpasses the performance of SVMs on the image data. Furthermore, classification performance increases with the amount of spatio-temporal

⁵Note that the very similar matching algorithms presented in section 5.2 achieved better performance - nevertheless, the *relative* performance in comparison with SVMs is important in this context.

Figure 7.4: Recognition performance of the KNN classifier for $\mathcal{M}_{pf,pfv}$ as a function of a) the geometric parameter σ_{dist} and b) of the number of maximum features Q used in the matching process.



KNN 2-nearest-neighbor	recognition error
$M_f(F_{t_\alpha}, F_{u_\alpha})$	17.50 %
$M_{fp}(F_{t_\alpha}, F_{u_\alpha})$	14.50 %
$M_{fpv}(F_{t_\alpha}, F_{u_\alpha})$	2.0 %

SVM 1-vs-rest classifier	optimal parameter	recognition error	# SVs
$M_f(F_{t_\alpha}, F_{u_\alpha})$	$\gamma = 12$	6.25 %	180.2
$M_{fp}(F_{t_\alpha}, F_{u_\alpha})$	$\gamma = 6$	3.50 %	115.7
$M_{fpv}(F_{t_\alpha}, F_{u_\alpha})$	$\gamma = 1.5$	0.00 %	64.4

Table 7.3: KNN and SVM recognition results on the various feature metrics.

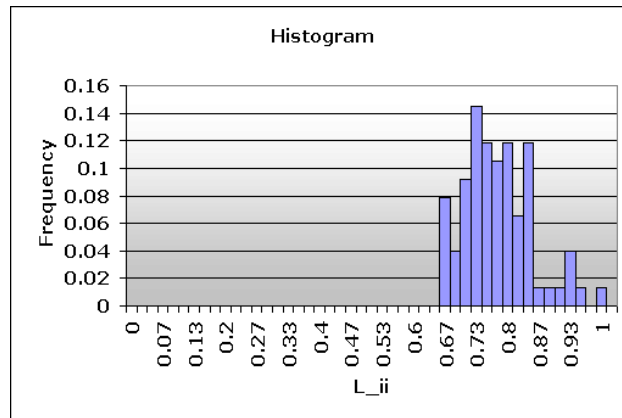
information added to the representation—with the best results obtained with the representation, which uses the direction vector as an explicit temporal prior. In addition, two post-hoc analyses shown in Figure 7.4 investigated the performance of the KNN classifier for the "spatial" similarity metric M_{pf} and the full spatio-temporal similarity metric M_{pfv} as a function of two parameters. First of all, recognition performance was analyzed as a function of the geometric bias parameter σ_{dist} , which specifies how feature transformations should be weighted in relation to appearance information. When using Euclidean distance only, M_{pf} yields the lowest recognition errors for small values of σ_{dist} , indicating that the information contained in the pixel distance improves performance only with low weights. This behavior is different when switching to the full spatio-temporal similarity metric, which uses the direction vector to introduce a biased evaluation of the Euclidean pixel distance. Here, increasing the relative weight of the geometric information results in an increase in performance, which saturates at 1.25% recognition error. It seems that this bias is crucial for finding correct correspondences over depth rotations, which might not be reflected in the appearance information alone.

The classification results for multi-class SVMs reported in Table 7.3 were obtained by integrating the similarity metrics into a RBF-kernel:

$$\exp(-\gamma(1 - M_{f,fp,fpv}(F_{t_\alpha}, F_{t_u})))$$

The most striking result is that using any various local feature representations results in a significant performance gain compared to the raw image representation. This mirrors the previous finding with K -NN classifiers and is further evidence for the robustness and salience of the chosen

Figure 7.5: Histogram of the diagonal values L_{ii} of the matrix \mathbf{L} , which was obtained through the Cholesky decomposition of the kernel matrix \mathbf{K} during a full SVM training run. Since the histogram contains only positive values this is proof that the matrix \mathbf{K} is positive definite.



spatio-temporal representation. In addition, we find that similarly to K -NN classification, performance increases with the amount of information used in the similarity metric. The full temporal similarity metric again yields the best results in this classification task and thus demonstrates the benefit of incorporating temporal information into both representation and matching of visual feature sets. Finally, as could be expected, SVMs again achieve a significant increase in classification performance compared to K -NN classification.

These results thus demonstrate that a local feature representation, which makes the spatio-temporal properties of the visual input data explicit, in combination with a robust classification scheme from machine learning results in an excellent classification performance.

7.2.4 Experimental validation of positive definiteness

Histogram of the diagonal values L_{ii} of the matrix \mathbf{L} , which was obtained through the Cholesky decomposition of the kernel matrix \mathbf{K} during a full SVM training run.

In order to be able to integrate the novel feature similarity metrics introduced in section 2 into kernel machines, they need to satisfy Mercer's theorem. This theorem states that the resulting kernel matrices have to be symmetric positive definite. In this section, we want to present evidence that the proposed metrics meet these criteria.

First of all, it can be seen from the definition of the similarity score \mathcal{M} that the evaluation of correspondences is restricted to Q maximum values, which makes \mathcal{M} symmetric. The second prerequisite for Mercer's theorem is that the kernel matrix be positive (semi-)definite. A full proof of the positive definiteness of the kernel matrix seems to be difficult due to the nonlinear functions (such as the \max function) involved in determining the similarity score. However, it is possible to provide an empirical validation of the positive definiteness under the typical experimental conditions of this study.

A possible test for positive definiteness would be to verify that the smallest eigenvalue of the kernel matrix, which is of the form $\mathbf{K} = K_{ij} = \exp(-\gamma(1 - M_{f,fp,fpv}(F_{t_\alpha}, F_{t_u})))$, is positive. However, a much more efficient test, which avoids the computationally intensive calculation of the eigenvalues, is to check whether the Cholesky decomposition of $\mathbf{K} = \mathbf{L} \cdot \mathbf{L}^T$ exists. Recalling that the entries of the matrix \mathbf{L} —often termed the square-root of \mathbf{K} —are defined as

$$L_{ii} = \left(K_{ii} - \sum_{k=1}^{i-1} L_{ik}^2 \right)^{\frac{1}{2}}, \quad L_{ji} = \frac{1}{L_{ii}} \left(K_{ij} - \sum_{k=1}^{i-1} L_{ik} L_{jk} \right)$$

the Cholesky decomposition entails that if $(K_{ii} - \sum_{k=1}^{i-1} L_{ik}^2) < 0$ for some i , then \mathbf{K} cannot be positive definite. Figure 7.5 shows a histogram of L_{ii} for all possible kernel matrices evaluated during a full training run using $\mathcal{M}_{f_{pv}}(F_{t_\alpha}, F_{u_\alpha})$ as the similarity metric. The maximum value of this metric is $\mathcal{M}_{f_{pv}}(F_{t_\alpha}, F_{u_\alpha}) = 1$, which is therefore also the maximum value of L_{ii} shown in the histogram. More importantly

$$\min(L_{ii}) = 0.664 \rightarrow L_{ii} > 0 \quad \forall i$$

which provides empirical validation that all kernel matrices are positive definite.

7.3 Conclusion

In this chapter I proposed various spatio-temporal representations for image sequences based on sets of local visual features extracted from keyframes. A novel metric was developed which assesses feature matching directly in the kernel function using appearance, position and temporal information of the features. This was achieved by replacing the standard scalar product by a feature matching score. In addition, empirical evidence has shown that the proposed metrics satisfy all necessary criteria in order to be integrated with kernel machine frameworks. Although the computational experiments in this chapter were done using Support Vector Machines as a particular type of kernel machine, the similarity metrics can of course be integrated into other types of kernel machines.

The classification results showed that all spatio-temporal feature representations significantly decreased the classification error compared to a simpler representation containing raw pixel data. More importantly, the lowest classification errors were obtained by using information about spatio-temporal feature transformations as developed in the keyframe framework. This demonstrates the benefit of exploiting knowledge about spatio-temporal patterns in image sequences. In addition, the integration of the proposed data representations into the Support Vector Machine framework allowed the combination state-of-the-art computer vision techniques with robust classification schemes from statistical learning theory.

Future studies need to be done in order to integrate more sophisticated temporal priors into the matching process, which - rather than encoding global *average* motions - are based on “motion fields” and might thus be able to provide better discrimination performance also for complex object motions.

Chapter 8

General conclusion and outlook

In this thesis, I have presented an integrative framework which combines methods and results from cognitive research, computer vision and machine learning. The key challenge of the proposed keyframe framework that was identified in chapter 1 in the context of established as well as new experimental results from psychophysical research, was to develop a *structured* recognition framework in which objects are represented in a *spatio-temporal* context.

I have demonstrated how the proposed keyframe framework that followed from these considerations is able to model view-based, structured object recognition, temporal continuity in object learning as well as temporal properties of object representations. Furthermore, computational experiments have shown how these psychophysically motivated properties can lead to increased robustness in learning and recognition of objects. Finally, another central contribution of this thesis has been to link the local, spatio-temporal feature representations introduced with the keyframe framework with state-of-the-art kernel classifiers from machine learning.

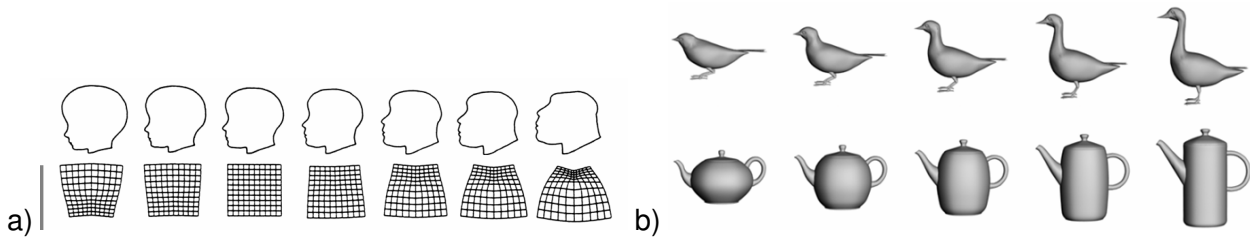
Again, it is important to stress that the aim of this thesis was not to implement a fully competitive object recognition system but rather to show how by adopting straightforward ideas from cognitive research, it becomes possible to make *conceptual* advances that help to

- come to a better understanding of the cognitive processes underlying object recognition and learning on a functional level [Marr, 1982]
- design "smarter", *cognitive* systems instead of "brute-force" computational recognition systems
- bridge the gap between different disciplines and show the synergies between them.

In order to show how one could further develop each of the three main foci (see Figure 1) of this thesis I want to present three studies that highlight possible extensions of the ideas presented so far. In particular, I will discuss

- theoretical integration of the keyframe approach with a more general *cognitive* framework with a particular focus on establishing a hierarchy of object recognition including processing strategies for categorization and context
- extension of the keyframe approach to integrate multiple modalities, such as vision and haptics in a perception-action context, that is, *multi-modal keyframes*
- computational experiments using the local kernel classifier that go beyond recognition of single exemplars towards recognition of *object categories*

Figure 8.1: a) Illustration of shape deformations to create a basic level category b) Examples of 3D stimuli used in the psychophysical experiments in Graf et al. [2002]



8.1 A unified framework for object recognition

In this section, I want to try and integrate the keyframe framework within the broader context of a "grand model" of visual object recognition (see Figure 8.2). This model was developed as part of the CogVis project which aimed to design and build successful cognitive vision systems inspired by cognitive research. The model summarizes a number of cognitive studies Graf et al. [2002] that aimed to investigate several issues in the field of object recognition and categorization.

More specifically, the questions addressed were

- the role of components or local information in object recognition (see chapter 1)
- how object motion and shape interact in a categorization task (see chapter 1)
- how shape similarity and basic level categorization can be accounted for
- whether context and top-down information is important during recognition

Since the first two issues were already covered in chapter 1, in the following, I will briefly summarize the results of the studies concerning the last two points. In addition, I want to discuss how one could provide links from the psychophysical results of these studies to the proposed keyframe framework and how to bring everything together into a unified framework for object recognition.

8.1.1 Categorization processing by feature correspondences

Two studies in Graf et al. [2002] dealt with the question how a model of basic level categorization should be conceptualised. The motivation behind these studies was the observation that different members of the same basic level category usually can be aligned by rather simple deforming *shape transformations* (see Figure 8.1a). Thus, shape variability within basic level categories could be *described* well with such transformations. The studies investigated whether categorization performance and perceived similarity are systematically related to these shape transformations.

Study 2 in Graf et al. [2002] showed that the perceived similarity of line drawings is systematically related to the amount of shape transformation. These results could not be explained by simple affine transformations or changes in the configuration of parts, because highly similar results were found also for those categories with little affine change, or with small changes in the configuration of parts. Study 3 extended this line of research by investigating categorization performance in a speeded categorization task using images of 3D objects (see Figure 8.1b). Experimental results confirmed the findings with line drawings by again showing that categorization performance deteriorated systematically with increased shape transformation. The same decrease was also found when objects were rotated in the image plane. Overall, the systematic dependency on the amount of shape transformation was demonstrated for rating tasks and speeded categorization

tasks, for line drawings and grey-level images, as well as for upright and plane rotated objects. Interestingly, when both transformational distance *and* image-plane orientation were manipulated, the two effects did not interact, which suggests that they were compensated independently.

The findings of these studies suggest an *image-based* model of categorization, which is in accordance with earlier findings (for example, Rosch et al. [1976], see also section 1.2.3). The systematic relation between categorization performance and the amount of shape transformation is reminiscent of the dependency of recognition performance on the amount of rotation and size-scaling. In principle, different image-based models may account for these results. In Graf et al. [2002], the authors argue for an *alignment* model of categorization, in which categorization is achieved by an image-based alignment of a memory representation with the stimulus representation. The proposed alignment model with deforming transformations implies that corresponding parts or features are identified and brought into correspondence, and thus entails again the notion of a structured representation. Therefore, an alignment that allows for deforming transformations is compatible with the notion of structured category representations, and - unlike the structural description approaches put forward by Biederman [1987] - does not imply the problematic notion of invariance to spatial transformations (see also chapter 1).

How does the keyframe approach fit into this notion of spatial transformations? Recall, that in Study 3 it was shown that effects of rotation and transformation were independent - in Graf et al. [2002] a possible neurophysiological model is developed based on the findings of Wang et al. [1998]. In this neurophysiological study, positions of activation spots changed gradually along the cortical surface as a stimulus face was rotated in depth. Interestingly, intermediate orientations were coded in intermediate locations on the cortical surface. Such an organization of views in the cortex would be highly compatible with the keyframe approach which also is based on clustering of similar views. Indeed, when an object in an unusual orientation has to be recognized, the activation could proceed along these cortical paths - in a similar manner as in the linked keyframe representation - such that more time is required for increasing orientation differences. This concept might then also be transferred to representations of *categories* with a topology created by transformational distances between shapes. Similar to the recognition process, matching of similar shapes may proceed along these neuronal pathways, corresponding to a time-consuming compensation process. In accordance with the results from Study 3, compensation processes for orientation and shape could then be processed by different modules in the cortex. This neurophysiological study therefore could provide evidence on how to enable both recognition and categorization by the *same functional modules* in the brain, which links the keyframe framework with the transformational processes needed for categorization.

However, one could go one step further: As was shown in chapter 3, keyframes explicitly encode information about feature transformations between linked keyframes. Thus it might be possible to capture the transformational similarities between different examples of a category by these feature transformations. Indeed, as both the examples on keyframe extraction with morphed and scrambled sequences (see chapter 4) and the examples of matching between two different faces in chapter 4 have shown, it seems possible to acquire category representations by means of local feature correspondences¹. Apart from fitting well into the psychophysical evidence presented here, the advantages of using the keyframe framework to model categorical transformations are:

- Explicit access to the *correspondence fields*: This could for example be used to generate *image-based* morphable models (see also Blanz and Vetter [1999]), which in addition to being able to recognize novel images could also be used to *synthesize* category exemplars.
- Learning of categories: learning is simply based on recognition of exemplars and can be integrated seamlessly with the online learning strategy. In addition, build-up of categorical representations (for basic level categories) should be automatic.

¹In principle, the morphable model framework introduced in Blanz and Vetter [1999] also rests on this assumption.

- Canonical exemplars: In the same manner as shown in chapters 1 and 4, viewing statistics can be used to create canonical exemplars of categories - this could for example be tested with the typicality ratings done in Graf [2002].

It seems that the keyframe framework presented in this thesis is well suited for modeling category processing as well; experiments using both artificial and real-world stimuli would be an interesting line of future research.

8.1.2 The role of context in object recognition

Another study in Graf et al. [2002] explored the role of context. The motivation for this study was the observation that objects in the real-world do not occur in a random manner (see also Figure 1.2). Looking for a chair in an office is more successful than looking for a chair outside as the correct *scene context* is present (see also Biederman et al. [1981]) - similarly, tea spoons tend to be found near tea cups, which could be termed an *object context*. A cognitive system, which is adapted to the environment, should take such *co-occurrences* into account and could use them in the form of priors for faster and less viewpoint-dependent recognition. This could be achieved by *priming* image-based representations in visual memory (top-down arrow and green area in top left of Figure 8.2). Such a priming would be especially helpful when *non-canonical* views have to be recognized, which often tend to be similar to views of other objects.

Indeed, a study in Graf et al. [2002] confirmed that non-canonical views of an object could be recognized much faster, when they were primed by another object that tends to co-occur in the same scene. In this study, several household objects were used as stimuli which were grouped according to different object contexts (examples for matching pairs from the same object context would be: cup/teapot, remote-control/TV). In addition, the objects were available in two different viewpoints - a canonical viewpoint and a non-canonical viewpoint. The psychophysical experiment followed a standard priming design in which a first image was presented, which was followed shortly after by a second - testing - image that had to be recognized. Each priming image could either be from the same or a different context whereas each testing image could be shown in either the canonical or non-canonical view. The most interesting effect observed was a clear interaction between context and viewpoint - more specifically, a consistent context would be a comparatively better priming stimulus for a non-canonical viewpoint than for a canonical viewpoint. This confirms the hypothesis stated above that contextual priming could be helpful in situations where difficult, that is, non-canonical viewpoints have to be recognized.

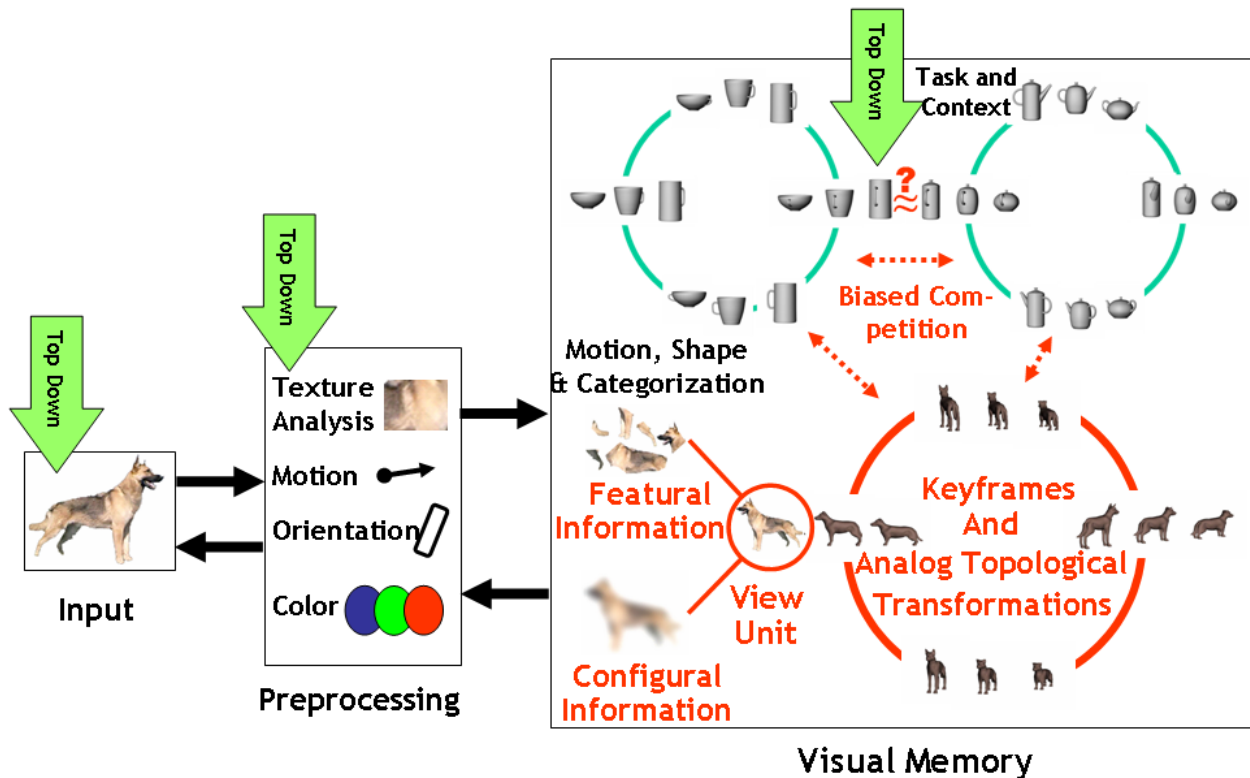
Integration with the keyframe approach for contextual processing seems fairly trivial. It would basically amount to connecting or linking keyframe representations for several objects based on the recognition statistics provided by each keyframe. In a suitably rich description of the visual world, which is represented by a network of keyframes, subnets of activation (that is, recognized keyframes) could then be used to define a visual context.

Another interesting way to integrate contextual processing of *scenes* into the keyframe framework could be along the lines of the approach proposed by Torralba and Oliva [2003], in which the statistics of natural scenes were investigated. Using simple histograms in Fourier space, which evaluated the frequency contents in several orientation bands, they were able to accurately recognize several categories of various outdoor and indoor scenes. This type of information could, for example, be used to determine a context prior, which is connected with the learned keyframe representations.

8.1.3 Summary

Figure 8.2 shows how one might integrate the various psychophysical studies into a unified framework of object recognition. In this figure (read from right to left), visual recognition and categorization is achieved by matching the visual input to stored memory representations. The visual input

Figure 8.2: Unified framework of object recognition. Shown are various components of a unified framework which models context effects and includes a categorization hierarchy based on spatio-temporal representations and structured processing.

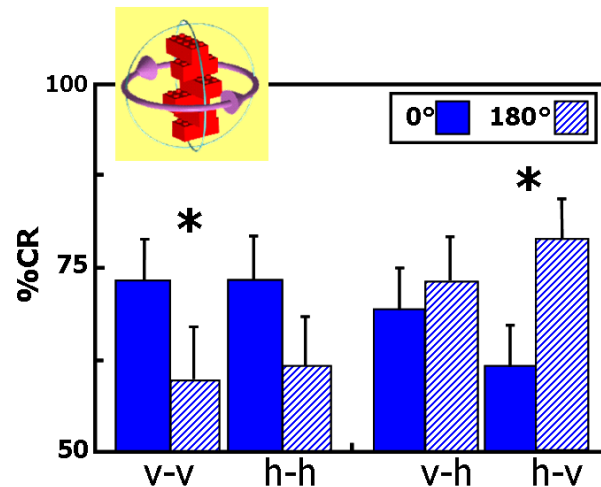


consists of a spatio-temporal flow of visual information, which is induced by the object or scene. In a preprocessing stage, low-level features like color, orientation, motion, texture and other properties are extracted from this input in an appearance-based manner. These features provide the basis for selecting possible candidate representations from visual memory in order to match them against the input for both recognition and categorization.

As was shown in chapter 1, matching of such visual representations can occur both on the basis of component and configural information. The outputs of component and configural matching are then pooled in view-units which basically induce a view-based input representation. The novelty of this model lies in the combination of categorization *and* recognition processes in such a view-based framework - this is possible because basic categorization processes still relies on image-based similarity, which can be mediated by a view-based appearance representation.

The spatio-temporal continuity of the visual input plays a fundamental role in the learning of such object representations (chapter 1), which represents a first interface to the spatio-temporal representations generated in the keyframe framework. As we have seen, recognition in the keyframe framework relies on bringing structured representations (for example consisting of local visual features) into correspondence. Going one step further, *basic level categorization* could be achieved by a similar alignment process, which brings similar object parts of the stimulus representation and memory representation into correspondence - experimental evidence for such a process was discussed above. In addition, studies on the saliency of spatial and temporal cues

Figure 8.3: Results (adapted) from the psychophysical experiments of Newell et al. [2001] on viewpoint dependence and cross-modal integration effects of visuo-haptic object recognition.



in a categorization task (see chapter 1) have shown the importance of spatio-temporal processes in categorization, which supports the idea of a *common spatio-temporal representation* for both recognition and categorization processes in the form of the keyframe framework.

Finally, the last important functional ingredient of the "grand model" is that all processes can be modulated by expectations, which are for example provided by the (scene) context. In the study discussed earlier, it was found that especially *unfamiliar* views of objects are facilitated by contextual information. This suggests that top-down processing in the form of *context information* could for example help to disambiguate two otherwise similarly looking images of two categories or objects.

This unified framework for object recognition confirms and extends existing image-based models of recognition and categorization along four lines: First, recognition involves both component and relational information. Second, similarity ratings and basic level categorization performance are systematically related to shape variations within basic level categories. Third, top-down contextual expectations play a role in object categorization. Fourth, motion cues can be integrated with form cues for categorization decisions. In the context of this framework, I have presented extensions of the keyframe approach, which could be used to integrate keyframes to enable categorization and context processing and which represent starting points for future research.

8.2 Multi-modal keyframes

Up to now, the keyframe framework has been treated as a framework for recognition of objects in the *visual* modality. The general idea of spatio-temporal object representations, however, could of course be extended to other modalities as well. In the following, I will introduce such a multi-modal object representation combining visual with proprioceptive information, which was successfully implemented on a robot-setup and subsequently tested in object learning and recognition scenarios.

8.2.1 Psychophysics of visuo-haptic object recognition

As mentioned in chapter 2, recent research in neuroscience has led to a paradigm shift from cleanly separable processing streams for each modality towards a more integrative picture of networks of task-specific modules, each of which have access to multi-modal data representations.

Cross-modal integration of data from different modalities was also shown, for example, to play an important role for haptic and visual modalities during object recognition. In a recent psychophysical experiment (see Newell et al. [2001]), participants had to learn views of four simple, 3D objects made of stacked LEGOTM bricks either (see inset of Figure 8.3) through the haptic modality (when they were blind-folded) or through the visual modality (without being able to touch them). Testing was then done using an old-new recognition paradigm with four different conditions: two within-modality conditions, in which participants were trained and tested in either the haptic or the visual domain and two between-modality conditions, in which information from the learned modality had to be transferred to the other modality in order to solve the recognition task. For each condition, in addition, either the same viewpoint as in the learned condition was presented or a viewpoint rotated by 180° around the vertical axis, such that viewpoint-dependency of object recognition could be tested as well. The recognition results for the four conditions shown in Figure 8.3 (percent correct responses, %CR) demonstrate first of all that cross-modal recognition is possible ("v-h" and "h-v") well above chance (which would be at 25% for this experiment). Not surprisingly, recognition of rotated objects in the within-modality condition is severely affected by rotation in both modalities. This shows that not only visual recognition is highly view-dependent but also that *haptic* recognition performance is directly affected by different viewing parameters. One could thus extend the concept of view-based representations of objects also to the haptic modality.

Another interesting finding of this study is that recognition performance in the haptic-to-visual condition *increased* with rotation. The authors assumed that this was an example of a true cross-modal transfer effect - the reason for such a transfer lies in the fact that during learning the haptic information extracted by participants was mainly coming from the *back* of the object. When presented with a *rotated* object in the visual modality, this haptic information was now visible, which enabled easier recognition.

The results from this experiment thus support the view that haptic recognition is also mediated by view-based processes - although the exact dependence on viewing angle is still a matter of investigation. In addition they shed light on how information from the haptic modality can be used to enable easier recognition in the visual modality (for a study on how visual and haptic cues are combined in the visual system, see also Ernst and Banks [2002]). Taken together with the spatio-temporal framework outlined in this thesis, this cross-modal transfer might be an important reason for the excellent visual performance of human object recognition - after all, it is known that infants learn extensively by grasping and touching objects, which thus could provide a "database" of object representations for visual recognition.

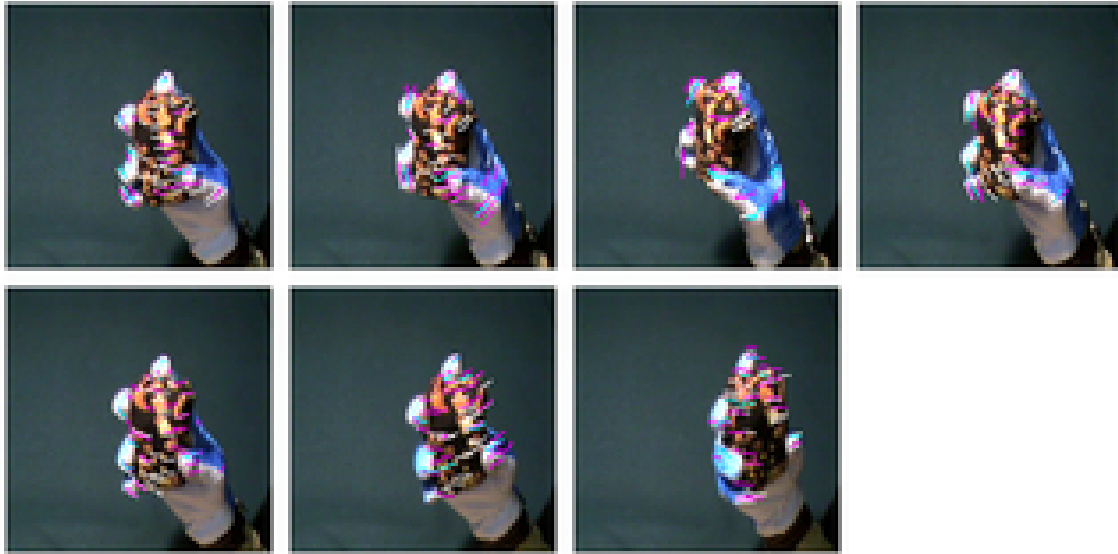
8.2.2 Multi-modal keyframes - the view transition map²

In the following, I want to describe how one can combine visual with proprioceptive input to create a multi-modal keyframe representation as well as a set of computational experiments which demonstrate the effectiveness of this approach.

A first example of the type of input that could be used to construct a multi-modal representation is depicted in Figure 8.4, in which the keyframe analysis was applied to sequences coming from a recording setup at DIST, Genoa. These video sequences show a hand manipulating an object from the viewpoint of the human who is exploring the object manually. With a typical set of parameters, the keyframe system found seven keyframes (for a total sequence length of 72 frames), which are shown in Figure 8.4 together with their feature trajectories. The interesting point here is that for this sequence not only data from the visual modality was recorded but also *proprioceptive* data, which specifies the rotation of the wrist as well as the shape of the hands while manipulating the object. Taken together, such a representation enables to learn *multi-modal view-based representations* of

²The multi-modal representation as well as the experiments described were developed in collaboration with Sajit Rao and Lorenzo Natale at the Dipartimento di Informatica, Sistemistica e Telematica at the University of Genoa.

Figure 8.4: Keyframes from a sequence showing manual exploration of an object that consisted of turning movements. Note the strong rotation components visible in the feature trajectories (shown as overlays).



objects in a straightforward manner as an extension of the keyframe framework. In the following, I want to provide some ideas of how such a multi-modal representation could be used to recognize objects more robustly.

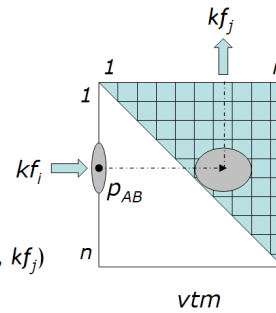
First of all, proprioceptive information about the 3D configuration of the hand could actually be used in a similar manner as in the psychophysical experiment mentioned in the previous section. This essentially 3D information can for example generate a 3D viewing space in which keyframes can be placed - anchored - at certain proprioceptive coordinates. This would link the visual appearance from the keyframe with the hand position and configuration and thus provide a *proprioceptively anchored visual space*. Returning to Figure 3.1, we see that one of the inherent disadvantages of the keyframe framework is that the *real-world topology* of the keyframe graph is undefined - only the outgoing and incoming links for each keyframe are known. Although this provides enough information to solve recognition tasks (see chapter 6), being able to convert the viewer-centred keyframe graph into an *object-centred* keyframe graph would provide additional constraints for matching visual appearances as such a representation is much closer integrated into a perception-action loop.

One of the problems with the idea of a proprioceptive space, however, is that absolute coordinates in such a space make little sense from a recognition point of view. Consider, for example, a hand picking up and manipulating an object in order to investigate it and build a visual keyframe representation; although it might be the case that objects suggest a *canonical grasp* (in much the same manner as it might suggest an *affordance* in the Gibsonian sense), usually it is possible to pick up and hold an object in a number of ways - all of which will change the absolute proprioceptive coordinates to which keyframes will be attached. In order to overcome this limitation, I want to suggest to interpret the proprioceptive space in a similar manner to the keyframe graph: as a representation, which is based on *changes or transitions* in its underlying modality. Thus, rather than providing an absolute frame of reference, each generated keyframe could be attached to a relative change in proprioceptive coordinates. One way to implement such a *view transition map* is as a lookup table or matrix, in which each entry is accessed by its *transition* in proprioceptive space - this transition can consist of the Euclidean distance between two proprioceptive states vector of the hand (such as wrist angles, finger positions, etc.).

Figure 8.5: Match algorithm for the view transition map

Given: a new transition consisting of two keyframes and one proprioceptive transition (kf_A, kf_B, p_{AB}) & a view-transition-map (vtm)

- Find view matches kf_j for kf_A in vtm
- Use proprioceptive transition p_{AB} from kf_i to select possible candidate keyframes kf_j
- Match all kf_j with kf_B using view matching
- $Matchscore_{jB} = ViewMatch(kf_A, kf_i) \cdot ViewMatch(kf_B, kf_j)$
- Select the best such match



In order to illustrate the usefulness of such a representation in a recognition task, let us consider the following situation: First of all, a keyframe representation of an object is learned in an active exploration stage using a pre-learned motor program, which for example grasps an object and turns it around. Following the algorithm outlined in chapter 4, the start of the sequence marks the first keyframe kf_A . As soon as the second keyframe kf_B is found through visual tracking, the proprioceptive transition p_{AB} between the two keyframes is calculated. A 2x2 matrix is allocated in which $A_{1,2} = -A_{2,1} = p_{AB}$ and $A_{1,1} = A_{2,2} = 0$. With the third keyframe kf_C , the matrix grows to 3x3 dimensions where $A_{1,3} = A_{1,2} + A_{2,3}$ and so on until the end of the motor program, which also signals the end of the exploration or training phase. In a second step, a test object is picked up and keyframes are extracted again while the same motor program is executed. In order to recognize this object using the transition map (see Figure 8.5), the first keyframe that was generated from the test sequence is matched against all of the keyframes of the training sequence using visual similarity only (see chapter 4). Once this match has been established, the transition map can be used to quickly find neighboring keyframes by looking for the most similar proprioceptive transition from that keyframe that matches the current change in the proprioceptive state. These candidate keyframes are then compared to the second keyframe of the test sequence with the final matchscore consisting of the product of the two visual matches. With this strategy, one could expect to recognize objects in a much more efficient manner as the indexing into proprioceptive transitions allows for direct matches in an object-centered reference frame.

8.2.3 Computational experiments

In the following, I want to demonstrate the usefulness of the proposed representation using two computational experiments on self-terminating learning of objects and on recognition of several previously learned objects.

Experimental Setup: Figure 8.6a) shows the robot setup that was used for the experiments. The most important components of the setup in the context of this study consist are:

- an actively foveating stereo camera head that uses space-variant image sensors mimicking the distribution of rods and cones on the human retina - the computational experiments, however, used only the central, fovea part of the cameras with an image resolution of 128x128 pixels
- an anthropomorphic robotic arm with a hand (see Figure 8.6b+c) that has 6 actively controlled degrees of freedom and is fitted with 20 touch sensors.

The camera head was pre-programmed to fixate on the location of the hand in order to track the hand during all movements. In addition, a trajectory for the hand movement was defined, which

Figure 8.6: The robot setup that was used in the multi-modal keyframe experiments. a) full view of the setup b) close-up of the hand with touch sensors visible c) hand grasping an object

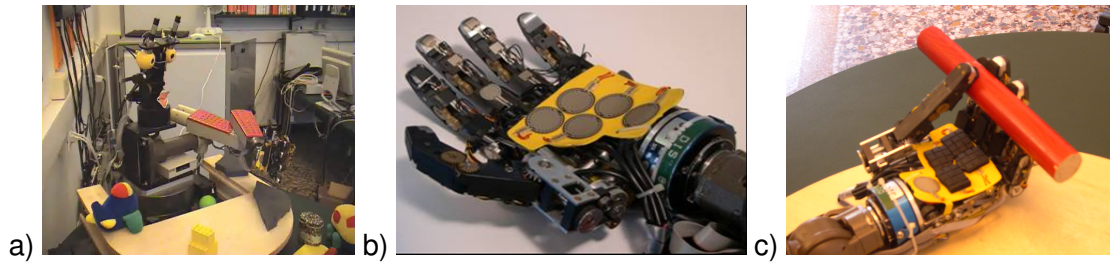


Figure 8.7: Objects used in the learning and recognition experiments.



consisted of the hand rotating first around the axis defined by the arm ("turning the hand") and then around a second axis resulting in an up-and-down movement of the hand. This exploratory motion sequence ensured an adequate visual coverage of any grasped object and can be seen as an approximation of manual exploration of objects by a human observer.

The test objects for the experiments consisted of 9 household and toy objects and are depicted in Figure 8.7 together with 2 control "objects" which simply consisted of the empty hand of the robot. Note that some of the objects are rather similar in terms of their shape.

Experiment 1 - Self-terminating object learning: In the first experiment, an object was placed into the robot's hand and the exploratory motion sequence was initiated. The visual input from the foveated cameras was then used to track features in real-time using a modified version of the keyframe framework based on the Kanade-Lukas-Tomasi tracker (Tomasi and Kanade [1991]). Each time the system found a keyframe, the proprioceptive transition leading from the last to this current keyframe was stored and used as an index into a matrix where each entry stored the visual information (in this case simply consisting of the whole frame rather than its local feature representation). In addition, each incoming keyframe was matched against all existing keyframes using the matching procedure outlined above (see Figure 8.5). If a match of suitable strength was found, the keyframe was discarded (similarly to the incremental learning procedure tested in section 5.3) otherwise the keyframe was inserted into the representation.

Figure 8.9 shows the results for three different objects. Taking the box object (Figure 8.9b) as an example, the full exploration sequence results in a total of 90 keyframes, which are presented to the keyframe system one after the other. For the first 10 keyframes, each incoming keyframe is added to the object representation. Subsequent keyframes, however, are a result of the repeated

exploration sequence and only rarely add new information to the representation. Consequently, the amount of *predicted* views approaches a diagonal and with this the object representation saturates around the 50th keyframe. The same pattern holds true for the other two objects - even for a shorter exploration sequence as shown for the brick object. This experiment confirms earlier results on the keyframe representation and represents a first test for the suitability and robustness of the multi-modal matching process in the context of object learning.

Experiment 2 - Multi-modal recognition of objects: In the second experiment, 9 objects were learned in the manner described in Experiment 1. In addition to these objects, two *empty* sequences in which the empty hand executed the motion sequence were recorded as control conditions. To test the recognition performance of the learned view transition maps, 6 of the objects were again shown to the robot and the same movements were executed. Each new keyframe was then compared against all learned transition maps as described above (see Figure 8.5) and the amount of matches above a pre-defined threshold in each of the 11 transition maps was added up to a final matching score. If the sequence would be identical, all keyframes would be found in the map and therefore the matching score would be 1. In order to provide a baseline as well as to be able to judge the influence of the addition proprioceptive information, a *visual-only* matching was also run in addition to the multi-modal matching procedure.

Figure 8.9 shows histograms of the matching scores for the two matching procedures (VTM and Visual Only) for the 6 test objects. Summarizing the overall recognition performance, using a simple winner-takes-all matching strategy, the view transition map is able to recognize 5 out of 6 objects correctly, whereas visual matching only recognizes 4 objects correctly. Although there is considerable variation in the absolute range of matching values, the VTM condition results in a much increased discriminability (such as for the Box, Gun and Toy objects) compared to the Visual Only condition. This is especially noticeable for the Toy object where it seems to be the extra information of the proprioceptive channel that enables correct recognition.

Summary: Although the two experiments presented here should not be seen as more than initial explorations into multi-modal object representations, the results are already very promising. Through a straightforward extension of the keyframe approach to include proprioceptive information, we have shown how multi-modal object representations can be learned in an unsupervised manner (in real-time) as well as how such representations can help to increase the discriminability of object recognition. The three most important performance improvements for the framework should be to

- integrate more sophisticated local feature matching instead of the holistic appearance matching done in these experiments,
- integrate robust classification schemes such as done in chapters 6 and 7,
- evaluate different cue combination approaches in order to optimally use proprioceptive and visual information.

Another interesting application for the transition map could also be used to execute specific motor actions based on visual input. Consider, for example, a situation in which an object has to be manipulated in order to insert it into a slot. The *inverse* of the transition map would allow such a task to be solved by executing the motor commands which trace out a valid motor path to the desired view based on the current view. In a similar manner, the transition map could also be used for efficient imitation learning based on visual input and for executing mental rotations. The key to all of these applications is that the transition map provides a strong coupling between proprioceptive data (action) and visual data (perception) and in this manner enables to represent a *perception-action loop* in an effective and efficient way.

Figure 8.8: Example results from the self-terminating learning experiment for a) the bottle object, b) the box object, c) the bricks object. As new keyframes are added to the representation (x-Axis), the amount of predicted views by the existing model grows.

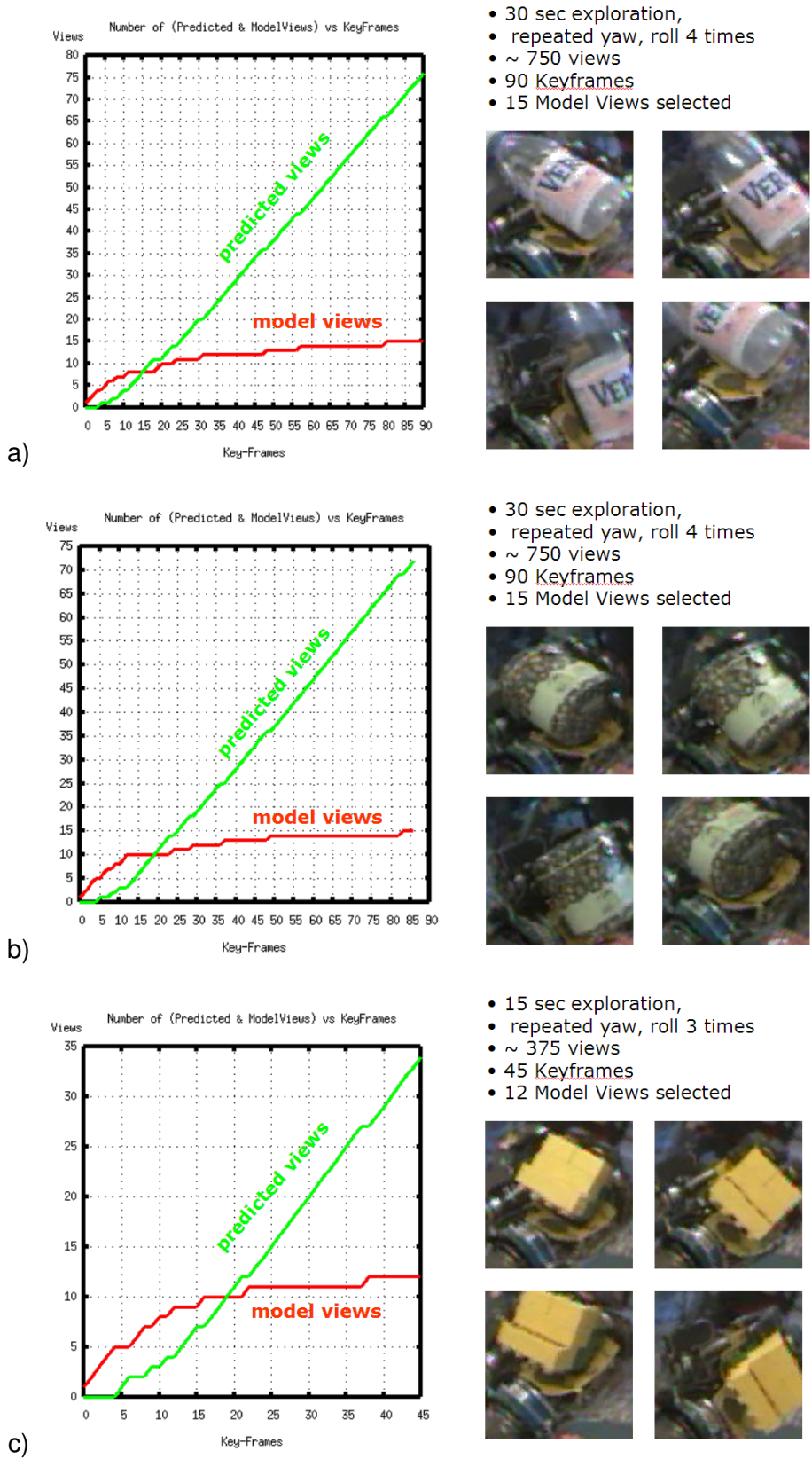
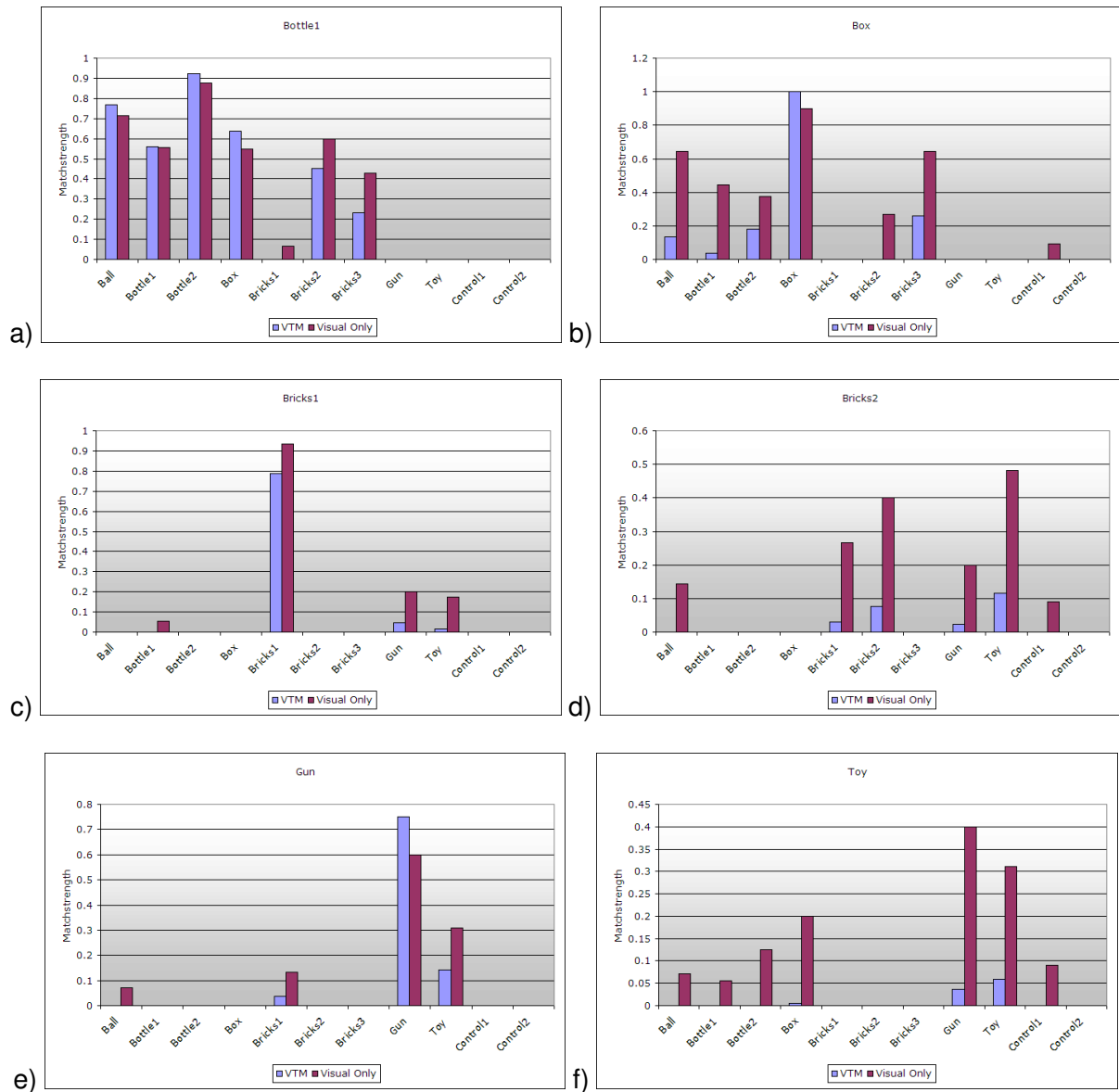


Figure 8.9: Results from the object recognition experiment for the objects: a) Bottle, b) Box, c) Bricks1, d) Bricks2, e) Gun, f) Toy. The light bars show matching using the view transition map, the dark bars show visual matching only.



8.3 Categorization using SVMs and local features³

In chapters 6 and 7 a framework for integration of support vector machines and local features for robust recognition of objects was proposed. Here, I want to present computational experiments in which it was tested whether a straightforward extension of the framework might also be suitable for *categorization* of objects (see also section 8.1 for a discussion of the importance of integrating categorization into computer vision approaches).

8.3.1 Experiment 1 - Categorization using a controlled database

The first experiment was done to evaluate the performance of the local feature framework in categorizing object views taken in a controlled setting. For this purpose, the CogVis-ETH80 database was used (see chapter 6), which contains a total of 80 objects from 8 different categories (apple, tomato, pear, toy-cow, toy-horse, toy-dogs, toy-cars and cups) shown on a homogeneous background. From this database, 16 evenly-spaced views for each object were selected. For each view we extracted one local feature representation (jet features Schmid and Mohr [1997], see also previous chapter), as well as two globally evaluated appearance-based representations (RGB histograms Swain and Ballard [1991] with 16 dimensions per color channel, or Gaussian derivatives Schiele and Crowley [2000] with three different filter kernel sizes $\sigma_{1,2,3} = \{1, 2, 4\}$). As in the previous chapter, we compared the performance of the local feature kernel with a standard nearest neighbor classifier (NNC).

For each of the 8 different categories, the 10 objects were divided into a training and test set, where the number of training objects was varied in the experiments in order to test the *generalization capabilities* of the local feature framework. Experiments were done with 1, 5 and 9 objects in the training set, with the remaining objects (9,5,1) in the test set. In order to increase statistical significance, each experiment was performed on 10 different partitions of the database with the final results being the average of the 10 runs.

Regardless of the size of the training set, the results in Figure 8.10 show that SVM + jet features consistently achieve the best performance in this categorization task followed by NNC + jet features. This result confirms the effectiveness of local features for categorization (in this controlled setting). It also underlines that the performance of the proposed local feature method is not due only to the representations used, but is made possible by their integration into the SVM framework.

Categorization results obtained with 9 objects in the training set can be compared with those reported in Leibe and Schiele [2003] who investigated categorization performance of different single-cue and multi-cue approaches using the same database. The best result using a (rather elaborate) single cue as well as nearest-neighbor classification reported in Leibe and Schiele [2003] is **86.40%** recognition rate, whereas the combination of SVM and straightforward local features yields **90.25%** recognition rate.

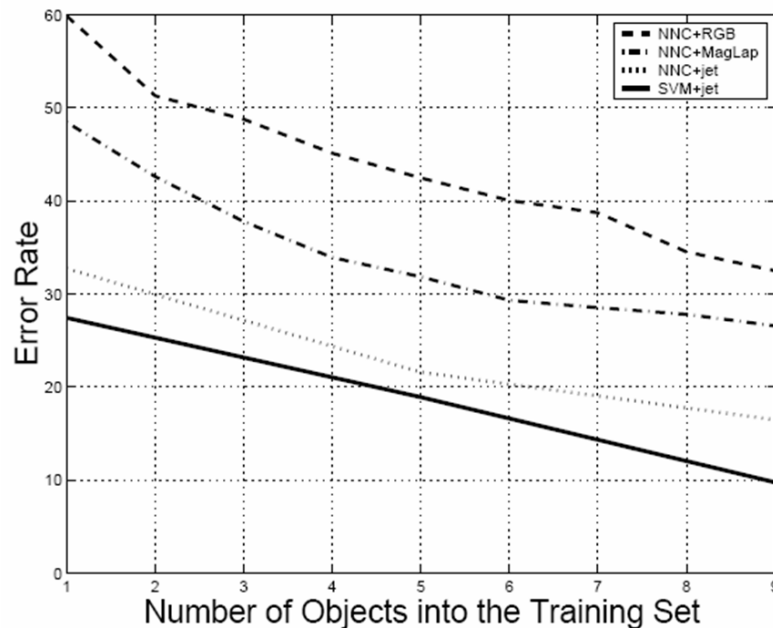
8.3.2 Categorization experiments in cluttered scenes

As already stated previously, the real challenge for computer vision is to recognize object categories in real-world settings under a variety of viewing conditions. In order to test whether the proposed framework of local features might be suitable for use in a real-world context, two further categorization experiments were performed.

In Experiment 2, training was done on object views taken in a controlled setting as in the previous section, whereas testing was done on a collection of pictures of the same category of objects *found on the web*. This experiment thus investigates the capability of the framework to generalize from "toy-objects" to real objects. In Experiment 3, both training and testing were

³This work was done in collaboration with B. Caputo from KTH Sweden and is published as Caputo et al. [2004].

Figure 8.10: Categorization performance of the local feature framework as a function of training set size.



performed on images collected under natural viewing conditions. The purpose of this experiment was to test the capability of the local feature framework to learn and recognize categories in cluttered views.

Experiment 2 - From toy to real-world objects: In this experiment, three categories out of the eight used in the previous experiment were considered due to the ready availability of image databases for testing - these categories were cars, cows and cups. Training was done on 10 objects from each of the cars and cows categories using the full set of 16 views per object. Testing was then performed on 135 views of real cars and 104 views of cows (see Figure 8.11 for examples). These two testing databases contained moderate variation in pose as well as in scale. In addition, we used the 82 images of the cup category as distractors during *testing* in order to investigate the amount of false alarms produced by each classifier.

In this experiment, SVM+jet features were benchmarked against NNC+jet features as de-

Figure 8.11: Examples of real-world images of cars and cows used in Experiment 2.

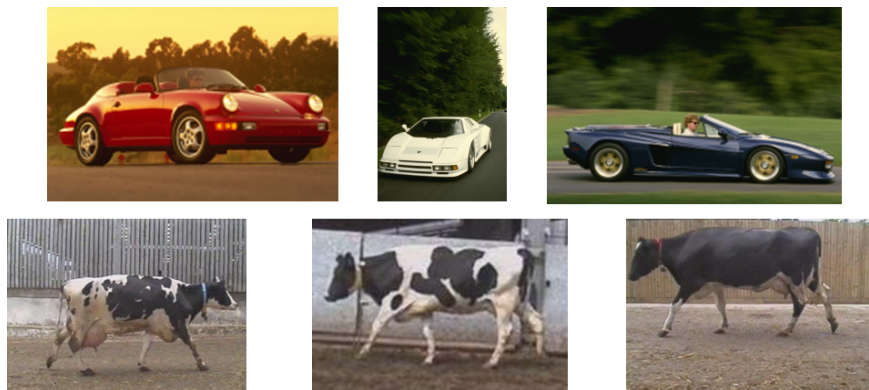


Table 8.1: Categorization performance for SVM (left) and NNC (right) classifiers and jet features. Training was done on views in homogeneous background and in the case of cars and cows, on toy-objects. Testing was performed on views of real objects taken in real-world settings.

SVM			NNC				
	cars	cows	cups		cars	cows	cups
car	86.1	11.7	2.2	car	78.1	17.5	4.4
cow	1.0	97.1	1.9	cup	2.9	90.4	6.7

Figure 8.12: Examples of real-world images of leaves, faces and cars as used in Experiment 3.



scribed in the previous section. The results reported in Table 8.1 demonstrate a clear advantage of the local kernel compared to nearest neighbor classification with both the amount of false alarms as well as misses drastically reduced. Although this experiment used only two categories in addition to a distractor category, the generalization performance seems already quite remarkable considering that training was done on toy objects only and that the database of test pictures consisted of real-world images,

Experiment 3 - Categorization in cluttered scenes: In this experiment, training and testing on real-world images was tested using 3 object categories from the Caltech database (see Weber et al. [2000a,b]) - cars (rear), leaves and faces. Each training and testing set consisted of 400 images for cars, 93 images for leaves and 218 images for faces. All of the images were taken in a real-world context and contained objects at different scales and under varying illumination conditions but without much viewpoint variation (see Figure for examples).

The training and testing protocol used followed those reported in Weber et al. [2000a,b] where the classifier was trained for *category detection* against a background class containing no objects. Training and test set for background consisted of the same number of views as the current category. Table 8.2 reports the results obtained with SVM+jet features as well as the results obtained with the probabilistic method as described in Weber et al. [2000a,b]. The local kernel algorithm, which integrates local features into a SVM framework, significantly outperforms the probabilistic method for all three categories. It is interesting to see that the performance pattern differs for the

Table 8.2: Categorization performance for SVM+jet features (left) and the probabilistic method proposed in Weber et al. [2000a,b] (right).

SVM+jet features			Probabilistic method		
cars	faces	leaves	cars	faces	leaves
97.88%	92.4%	91%	84%	87%	84%

two approaches: whereas the probabilistic approach achieves its best performance with the face category, the local kernel performs best on the face database. This could be due to the fact that the probabilistic method relies on a statistical description of parts and their geometric relations similar to the one given in chapter 1 - a description which was shown in chapters 1 and 4 to be particularly suited to the category of faces as these contain much stronger structural similarities than the other two categories. In contrast to this, the local kernel approach has so far not yet incorporated structural constraints which could explain the different performance pattern observed.

8.3.3 Summary

It seems that SVM combined with jet features, via local kernels, not only enables robust *recognition* of objects (see chapters 6 and 7), but can also be an effective approach for multi-object *categorization*. Future work will have to investigate, whether the proposed framework scales to a much larger number of categories (here, only 2 and 3 categories have been used for training and testing). As a larger number of categories will inevitably result in a much higher probability of single feature mismatches (see also chapters 6 and 7), integration of a suitable *position constraint* will be necessary to guarantee equally robust performance as in these preliminary experiments. This already points in the direction of the keyframe extensions discussed in chapter 5, which aimed at modeling the transformations of features between category members - such transformations could be one form of position constraint.

Acknowledgments

First of all, I want to sincerely thank Prof. Hanns Ruder for supervising this oeuvre. This thesis could not have flourished without the incredible support and intellectual stimulation of Prof. Heinrich Bülthoff at his department in the Max Planck Institute for Biological Cybernetics. I tremendously enjoyed its excellent, highly interdisciplinary working atmosphere as well as the unique opportunities that this place offered - both in terms of being able to meet the right people at the right places as well as to delve into a large variety of highly interesting and challenging topics (for which an expert was always close at hand). I especially want to thank all of my scientific collaborators with whom I had the fortune to work and publish - these are in alphabetical order: Martin Breidt, Heinrich Bülthoff, Barbara Caputo, Douglas Cunningham, Arnulf Graf, Markus Graf, Susanne Huber, Mario Kleiner, Lorenzo Natale, Fiona Newell, Manfred Nusseck, Sajit Rao, Adrian Schwaninger, and Sandra Schumacher. Special thanks go to all of the scientists (and reviewers) who worked in the CogVis-proposal on developing a cognitive vision system - the meetings, workshops and conferences have provided a great deal of inspiration for this work. Finally, throughout the whole process of designing, experimenting, and writing, my wife Miriam (who worked on her PhD-thesis at the same time) has been a never-ceasing source of inspiration, energy and love - a source for which I am eternally grateful.

Bibliography

- C. Altmann, H. Bühlhoff, and Z. Kourtzi. Perceptual organization of local elements into global shapes in the human visual cortex. *Curr Biol*, 13(4):342–349, Feb 2003.
- H.P. Bahrick, P.O. Bahrick, and R.P. Wittlinger. Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General*, 104:54–75, 1975.
- C. Baker, M. Behrmann, and C. Olson. Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nat Neurosci*, 5(11):1210–1216, Nov 2002.
- A. Barla, F. Odone, and A. Verri. Image kernels. *Proceedings of ICPR workshop on SVM*, 2002.
- P. Beardsley, P. Torr, and A. Zisserman. 3d model acquisition from extended image sequences. *Proc. ECCV'96*, pages 683–695, 1996.
- S. Belongie, J. Malik, and J. Puchiza. Matching shapes. *Proceedings of ICCV'01*, 2001.
- O. Ben-Shahar and S. Zucker. Geometrical computations explain projection patterns of long-range horizontal connections in visual cortex. *Neural Comput*, 16(3):445–476, Mar 2004.
- I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, pages 115–147, 1987.
- I. Biederman and P.C. Gerhardstein. Recognizing depth-rotated objects: Evidence for 3d view-points invariance. *Journal of Experimental Psychology: Human perception and Performance*, pages 1162–1182, 1993.
- I. Biederman and P. Kalocsai. Neurocomputational bases of object and face recognition. *Philos Trans R Soc Lond B Biol Sci*, 352(1358):1203–1219, Aug 1997.
- I. Biederman, R. J. Mezzanotte, J. C. Rabinowitz, C. M. Francolini, and D. Plude. Detecting the unexpected in photointerpretation. *Hum Factors*, 23(2):153–164, Apr 1981.
- H. Bischof and A. Leonardis. View-based object representations using rbf networks. *Image and Vision Computing*, 19:619–629, 2001.
- V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. *Proc. Siggraph99*, pages 187–194, 1999.
- V. Blanz, T. Vetter, H.H. Bühlhoff, and M.J. Tarr. What object attributes determine canonical views? *Perception*, pages 575–599, 1999.
- V. Blanz, S. Romdhani, and T. Vetter. Face Identification across Different Poses and Illuminations with a 3D Morphable Model. *Conference on Automatic Face and Gesture Recognition*, pages 202–207, 2002.
- R. Brunelli and T. Poggio. Face recognition: Features vs. templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, 1993.

- J. Bullier. Integrated model of visual processing. *Brain Res Brain Res Rev*, 36(2–3):96–107, Oct 2001.
- H.H. Bülthoff and S. Edelman. Psychophysical support for a 2-d view interpolation theory of object recognition. *Proceedings of the National Academy of Science*, pages 60–64, 1992.
- H.H. Bülthoff, S. Edelman, and M.J. Tarr. How are three-dimensional objects represented in the brain? *Cerebral Cortex*, pages 247–260, 1995.
- H.H. Bülthoff, C. Wallraven, and A. Graf. View-based dynamic object recognition based on human perception. *In Proc. ICPR'02*, pages 768–776, 2002.
- I. Bülthoff and H.H. Bülthoff. *Analytic and Holistic Processes in the Perception of Faces, Objects, and Scenes*, chapter Image-based recognition of biological motion, scenes and objects., pages 146–176. New York: Oxford University Press, 2003.
- C. Burges. *Geometry and Invariance in Kernel based Methods*. MIT press, 1998.
- B. Caputo and G. Dorko. How to combine color and shape information for 3d object recognition: kernels do the trick. *Advances in Neural information processing systems*, 15, 2003.
- B. Caputo, C. Wallraven, and M. Nilsback. Object categorization via local kernels. *ICPR (2)*, pages 132–135, 2004.
- S. Carey and R. Diamond. From piecemeal to configurational representation of faces. *Science*, 195 (4275):312–314, 1977.
- O. Chapelle, P. Haffner, and V. Vapnik. SVMs for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, 1999.
- N. Cristianini and J.S. Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge UP, 2000.
- C. Cyr and B. Kimia. 3d object recognition using shape similarity-based aspect graph. *In Proc. ICCV'01*, 2001.
- R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley and Sons, 2001.
- S. Edelman. *Representation and recognition in vision*. MIT Press, 1999.
- S. Edelman and N. Intrator. A productive, systematic framework for the representation of visual structure. *Advances in neural information processing systems 13*, pages 10–16, 2001.
- S. Edelman and D. Weinshall. A self-organizing multiple-view representation of 3d objects. *Biological Cybernetics*, pages 209–219, 1991.
- M.O. Ernst and M.S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415:429–433, Jan 2002.
- O. Faugeras. *Three-Dimensional Computer Vision: a Geometric Viewpoint*. MIT Press, 1993.
- D. H. Foster and S. J. Gilson. Recognizing novel three-dimensional objects by summing signals from parts and views. *Proc R Soc Lond B Biol Sci*, 269(1503):1939–1947, Sep 2002.
- V. Goffaux, B. Hault, C. Michel, Q. Vuong, and B. Rossion. The respective role of low and high spatial frequencies in supporting configural and featural processing of faces. *Perception*, 34(1): 77–86, 2005.

- A. Graf, A. Smola, and S. Borer. Classification in a normalized feature space using support vector machines. *IEEE Transactions on Neural Networks*, 14(3):597–605, 2003.
- M. Graf. *Form, Space and Object. Geometrical Transformations in Object Recognition and Categorization*. Wissenschaftlicher Verlag Berlin, Berlin, 2002.
- M. Graf, A. Schwaninger, C. Wallraven, and H.H. Bülthoff. Cognitive basis for recognition and categorization (D1.2, CogVis). *CogVis*, 2002.
- C. Harris and M. Stephens. A combined corner and edge detector. *Proc. 4th Alvey Vision Conf.*, pages 189–192, 1988.
- W. G. Hayward and M. J. Tarr. Differing views on views: comments on Biederman and Bar (1999). *Vision Res*, 40(28):3895–3899, 2000.
- B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio. Categorization by learning and combining object parts. *Proceedings of NIPS'01*, 2001.
- H. Hill, P.G. Schyns, and S. Akamatsu. Information and viewpoint dependence in face recognition. *Cognition*, 62(2):201–222, 1997.
- E. Hjelmås and B.K. Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83:236–274, 2001.
- C. Hsu and C. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.
- D.H. Hubel and T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol*, 160:106–154, Jan 1962.
- J. E. Hummel. Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual Cognition*, 8:489–517, 2002.
- J. E. Hummel and I. Biederman. Dynamic binding in a neural network for shape recognition. *Psychol Rev*, 99(3):480–517, Jul 1992a.
- J.E. Hummel and I. Biederman. Dynamic binding in a neural network for shape recognition. *Psychological Review*, pages 480–517, 1992b.
- A. Hyvarinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Netw.*, 13(4–5):411–430, 2000. ISSN 0893–6080. doi: [http://dx.doi.org/10.1016/S0893-6080\(00\)00026-5](http://dx.doi.org/10.1016/S0893-6080(00)00026-5).
- T. James, G. Humphrey, J. Gati, R. Menon, and M. Goodale. Differential effects of viewpoint on object-driven activation in dorsal and ventral streams. *Neuron*, 35(4):793–801, Aug 2002.
- T. Joachims. Svmlight. Technical report, Cornell, 2002.
- N. Kanwisher, J. McDermott, and M.M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. of Neuroscience*, pages 4302–4311, 1997.
- L. Kaufman and W. Richards. Spontaneous fixation tendencies for visual forms. *Perception & Psychophysics*, pages 85–88, 1969.
- M. Kirby and L. Sirovich. Applications of the karhunen-loeve procedure for the characterisation of human faces. *IEEE: TPAMI*, pages 103–108, 1990.
- J. Koenderink and A. van Doorn. The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32:211–216, 1979.

- Z. Kourtzi and N. Kanwisher. Representation of perceived object shape by the human lateral occipital complex. *Science*, 293(5534):1506–1509, Aug 2001.
- Z. Kourtzi, H. Bühlhoff, M. Erb, and W. Grodd. Object-selective responses in the human motion area MT/MST. *Nat Neurosci*, 5(1):17–18, Jan 2002.
- G. Krieger, I. Rentschler, G. Hauske, K. Schill, and C. Zetzsche. Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision*, pages 201–214, 2000.
- I. Laptev and T. Lindeberg. Interest point detection and scale selection in space-time. *Proceedings of Scale-Space'03*, 2003.
- B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. *Proceedings of CVPR'03*, 2003.
- B. Li and R. Chellappa. Face verification through tracking facial features. *Journal of the Optical Society of America A*, 18(12):2969–2981, 2001.
- N.K. Logothetis, J. Pauls, H.H. Bühlhoff, and T. Poggio. View-dependent object recognition by monkeys. *Current Biology*, pages 401–414, 1994.
- D. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, pages 355–395, 1987.
- D. Lowe. Toward a computational model for object recognition in it cortex. *Proc. BMCV'00*, pages 20–31, 2000.
- D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
- D. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer, 1985.
- D. Lowe. Object recognition from local scale invariant features. *Proc. of ICCV'99*, 1999.
- B S Mak and A H Vera. The role of motion in children's categorization of objects. *Cognition*, 71(1):11–21, May 1999.
- S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE TPAMI*, pages 674–693, 1989.
- D. Marr. *Vision*. San Francisco: Freeman Publishers, 1982.
- D. Marr and H.K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London*, pages 269–294, 1978.
- A. Massad, B. Mertsching, and S. Schmalz. Combining multiple views and temporal associations for 3-d object recognition. *Proc. ECCV'98*, pages 699–715, 1998.
- D. Maurer, R. Le Grand, and C. Mondloch. The many faces of configural processing. *Trends in Cognitive Sciences*, 6(6):255–260, 2002.
- B. Mel. Seemore: Combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation*, pages 777–804, 1997.
- K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. *Proceedings of ECCV'02*, 2002.

- K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005. ISSN 0162-8828. doi: <http://dx.doi.org/10.1109/TPAMI.2005.188>.
- Y. Miyashita. Neural correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, pages 817–820, 1988.
- H. Murase and S.K. Nayar. Visual learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, pages 5–24, 1995.
- F. Newell, C. Wallraven, and S. Huber. The role of characteristic motion in object categorisation. *Journal of Vision*, 4:118–129, 2004.
- F.N. Newell, M.O. Ernst, B.S. Tjan, and H.H. Bühlhoff. Viewpoint dependence in visual and haptic object recognition. *Psychological Science*, 12 (1):37–42, 2001.
- H. Op de Beeck, J. Wagemans, and R. Vogels. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat Neurosci*, 4(12):1244–1252, Dec 2001.
- A. O’Toole, Y. Cheng, B. Ross, H. Wild, and P. Phillips. Face recognition algorithms as models of human face processing. *FG*, pages 552–557, 2000.
- A. O’Toole, D. Roark, and H. Abdi. Recognizing moving faces: a psychological and neural synthesis. *Trends in Cognitive Sciences*, 6(6):261–266, 2003.
- S. Palmer. *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge, MA, 1999.
- S. Palmer, E. Rosch, and P. Chase. Canonical perspective and the perception of objects. *Attention and Performance*, pages 135–151, 1981.
- P. Penev and J. Atick. Local feature analysis: a general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500, 1996.
- G. Peters, B. Zitova, and C. von der Malsburg. How to measure the pose robustness of object views. *Image and Vision Computing*, pages 249–256, 2002.
- P. Phillips. Support vector machines applied to face recognition. *Proceedings of NIPS’99*, 1999.
- P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (10):1090–1104, 2000.
- M. Pilu. A direct method for stereo correspondence based on singular value decomposition. *Proc. CVPR’97*, pages 261–266, 1997.
- T. Poggio and S. Edelman. A neural network that learns to recognize three-dimensional objects. *Nature*, pages 263–266, 1990.
- M. Pontil and A. Verri. Support vector machines for 3-d object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):637–646, 1998.
- R. Rensink, J. O’Regan, and J. Clark. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8:368–373, 1997.
- I. Richardson. *H.264 and MPEG-4 Video Compression*. Wiley, 2003.
- M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, pages 1019–1125, 1999.

- I. Rock. *Orientation and form*. Academic Press, 1973.
- I. Rock and J. DiVita. A case of viewer-centered object perception. *Cognitive Psychology*, pages 280–293, 1987.
- S. Romdhani, V. Blanz, and T. Vetter. Face identification by fitting a 3d morphable model using linear shape and texture error functions. *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*, pages 3–19, 2002.
- D. Roobaert, M. Zillich, and J. O. Eklundh. A pure learning approach to background invariant object recognition using pedagogical support vector learning. *Proceedings of CVPR'01*, 2001.
- E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, pages 382–439, 1976.
- D. Roth, M. Yang, and N. Ahuja. Learning to recognize three-dimensional objects. *Neural Computation*, 14(5):1071–1103, 2002.
- F. Schaffalitzky and A. Zissermann. Viewpoint invariant texture matching and wide baseline stereo. *Proceedings of ICCV'01*, 2001.
- B. Schiele and J. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, pages 31–50, 2000.
- C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE TPAMI*, pages 530–535, 1997.
- C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, pages 151–172, 2000.
- B. Schölkopf and A.J. Smola. *Learning with Kernels*. The MIT Press, 2002.
- A. Schwaninger, F. Mast, and H. Hecht. Mental rotation of facial components and configurations. *Proc. Psychonomic Society 41st Annual Meeting, New Orleans, USA*, 2000.
- A. Schwaninger, S.M. Collishaw, and J. Lobmaier. Role and interaction of featural and configural processing in face recognition. *Journal of Vision*, 2:602, 2002a.
- A. Schwaninger, J. Lobmaier, and S.M. Collishaw. Role of featural and configural information in familiar and unfamiliar face recognition. *Biologically Motivated Computer Vision*, 2525:643–650, 2002b.
- A. Schwaninger, C.C. Carbon, and H. Leder. *Development of face processing*, chapter Expert face processing: specialisation and constraints, pages 81–97. Hogrefe, Göttingen, 2003.
- G. Schwarzer and D. W. Massaro. Modeling face identification processing in children and adults. *J Exp Child Psychol*, 79(2):139–161, Jun 2001. Clinical Trial.
- P. G. Schyns, R. L. Goldstone, and J. P. Thibaut. The development of features in object concepts. *Behav Brain Sci*, 21(1):1–17, Feb 1998.
- G. Scott and H. Longuet-Higgins. An algorithm for associating the features of two images. *Proc. Royal Society of London*, pages 21–26, 1991.
- J.H. Searcy and J.C. Bartlett. Inversion and processing of component and spatial-relational information of faces. *Journal of Experimental Psychology: Human Perception and Performance*, 22(4):904–915, 1996.

- J. Sergent. An investigation into component and configurational processes underlying face recognition. *British Journal of Psychology*, 75(Pt2):221–242, 1984.
- N. Sigala, F. Gabbiani, and N. K. Logothetis. Visual categorization and object representation in monkeys and humans. *J Cogn Neurosci*, 14(2):187–198, Feb 2002.
- T. Sim, R. Sukthankar, M. Mullin, and S. Baluja. Memory-based face recognition for visitor identification. *FG*, pages 214–220, 2000.
- P. Sinha and T. Poggio. Role of learning in three-dimensional form perception. *Nature*, pages 460–463, 1996.
- J.V. Stone. Object recognition using spatio-temporal signatures. *Vision Research*, pages 947–951, 1998.
- J.V. Stone. Object recognition: View-specificity and motion-specificity. *Vision Research*, pages 4032–4044, 1999.
- S. Stringer and E. Rolls. Invariant object recognition in the visual system with novel views of 3d objects. *Neural Computation*, 14(11):2585–2596, 2002.
- M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision*, pages 11–32, 1991.
- M. Tarr and H.H. Bülthoff. *Object recognition in man, monkey, and machine*. MIT Press, 1998.
- M. J. Tarr, P. Williams, W. G. Hayward, and I. Gauthier. Three-dimensional object recognition is viewpoint dependent. *Nat Neurosci*, 1(4):275–277, Aug 1998.
- M.J. Tarr and H.H. Bülthoff. Is human object recognition better described by geon structural descriptions or by multiple views? comments on biederman and gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance*, pages 1494–1505, 1995.
- M.J. Tarr and S. Pinker. When does human object recognition use a viewer-centered reference frame? *Psychological Science*, pages 253–256, 1990.
- M.J. Tarr, H.H. Bülthoff, M. Zabinski, and V. Blanz. To what extent do unique parts influence recognition across changes in viewpoint? *Psychological Science*, 8:282, 1997.
- P. Thompson. Margaret thatcher – a new illusion. *Perception*, 9(4):483–484, 1980.
- C. Tomasi and T. Kanade. Detection and tracking of point features. Technical report, CMU, 1991.
- A. Torralba and A. Oliva. Statistics of natural image categories. *Network*, 14(3):391–412, Aug 2003.
- N.F. Troje and H.H. Bülthoff. Face recognition under varying pose: the role of texture and shape. *Vision Research*, pages 1761–1771, 1996.
- M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, pages 71–86, 1991.
- S. Ullman. *The Interpretation of Visual Motion*. Cambridge, USA: MIT Press, 1979.
- S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE: TPAMI*, pages 882–905, 1991.
- S. Ullman, M. Vidal-Naquet, and E. Sali. Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, pages 682–687, 2002.

- L.G. Ungerleider and M. Mishkin. *Analysis of visual behavior*, chapter Two cortical visual systems, pages 549–586. MIT Press, 1982.
- V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition, 2000.
- V. Vapnik. *Statistical learning theory*. Wiley and Sons, 1998.
- T. Vetter, T. Poggio, and H.H. Bülthoff. The importance of symmetry and virtual views in three-dimensional object recognition. *Current Biology*, 4:18–23, 1994.
- R. Vogels, I. Biederman, M. Bar, and A. Lorincz. Inferior temporal neurons show greater sensitivity to nonaccidental than to metric shape differences. *J Cogn Neurosci*, 13(4):444–453, May 2001.
- Q. Vuong and M. Tarr. Rotation direction affects object recognition. *Vision Res*, 44(14):1717–1730, 2004.
- G. Wallis and H.H. Bülthoff. Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences of the United States of America*, pages 4800–4804, 2001.
- G.M. Wallis. The role of object motion in forging long-term representations of objects. *Visual Cognition*, pages 233–247, 2002.
- G.M. Wallis. Temporal order in human object recognition learning. *Journal of Biological Systems*, pages 299–313, 1998.
- G.M. Wallis and H.H. Bülthoff. Learning to recognize objects. *Trends In Cognitive Sciences*, pages 22–31, 1999.
- C. Wallraven and H.H. Bülthoff. Automatic acquisition of exemplar-based representations for recognition from image sequences. *CVPR 2001 - Workshop on Models vs. Exemplars*, 2001a.
- C. Wallraven and H.H. Bülthoff. Acquiring robust representations for recognition from image sequences. *DAGM-Symposium München 2001*, pages 216–222, 2001b.
- C. Wallraven, F.N. Newell, and S.A. Huber. The role of dynamic object properties in categorisation. *ESCAP 2001*, 2001.
- C. Wallraven, A. Schwaninger, S. Schumacher, and H. Bülthoff. View-based recognition of faces in man and machine: Re-visiting inter-extra-ortho. *Biologically Motivated Computer Vision*, 2525: 651–660, 2002.
- C. Wallraven, B. Caputo, and A.B.A. Graf. Recognition with local features: the kernel recipe. *Proc. of ICCV 2003*, 2003.
- G. Wang, M. Tanifuji, and K. Tanaka. Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neuroscience Research*, pages 33–46, 1998.
- M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. *Proc. ECCV2000*, pages 18–32, 2000a.
- M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. *Proc. CVPR2000*, pages 101–108, 2000b.
- J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. *Proceedings of ESANN'99*, 1999.
- B. Weyrauch, B. Heisele, J. Huang, and V. Blanz. Component-based face recognition with 3d morphable models. *CVPR '04*, page 85, 2004.

- J. Wieghardt and C. von der Malsburg. Pose-independent object representation by 2-d views. *Biologically Motivated Computer Vision*, pages 276–285, 2000.
- L. Wiskott, J. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.
- R. Yin. Looking at upside-down faces. *Journal of Experimental Psychology*, pages 141–145, 1969.
- A.W. Young, D.J. Hellawell, and D.C. Hay. Configural information in face perception. *Perception*, 16(6):747–759, 1987.
- S. Zeki. Improbable areas in the visual brain. *Trends Neuroscience*, 26(1):23–26, Jan 2003.
- Y.-M. Zhong and K. S. Rockland. Inferior parietal lobule projections to anterior inferotemporal cortex (area TE) in macaque monkey. *Cereb Cortex*, 13(5):527–540, May 2003.

Curriculum Vitae

Personal

Name

Christian Wallraven

Born

27th of March 1974, Krefeld, Germany

Education

1993

Abitur at Friedrich-List-Gymnasium in Reutlingen (average: 1.1)

1993-1999

studied Physics at the Eberhard-Karls-Universität in Tübingen

Oct 1997 - Mar 1998

DAAD - Stipend for studies at the University of Edinburgh

July 1998 - July 1999

Diploma thesis on *Stereo Reconstruction of Faces using Class-Based Knowledge*

July 1999

Diploma of Physics (with distinction) from the Eberhard-Karls-Universität in Tübingen

June 2000 -

PhD thesis on *A computational recognition system grounded in perceptual research*

Research

November 2002

co-organizer for the second international workshop on *Biologically Motivated Computer Vision*

May 2001 - June 2004

local co-ordinator for EU-funded COGVIS research grant

Publications (related to this thesis)

Proceedings

Bülthoff, H.H., Lee, S.-W., Poggio, T. and C. Wallraven: Biologically Motivated Computer Vision. Second International Workshop, BMCV 2002, Tübingen, Germany, November 22-24, 2002. Lecture Notes in Computer Science (2525), 662, Springer, Berlin (2002)

Journal Articles

Schwaninger, A., C. Wallraven and H.H. Bülthoff: Computational Modeling of Face Recognition Based on Psychophysical Experiments. *Swiss Journal of Psychology* 63(3), 207-215, Verlag Hans Huber, Hofgrete AG, Bern (2004) [Note: ISSN 1421-0185 doi:10.1024/1421-0185.63.3.207]

Wallraven, C. and A.B.A. Graf: Image Classification with SVMs using Spatio-temporal Feature Representations. (submitted) (2004)

Wallraven, C., A. Schwaninger and H. H. Bülthoff: Learning from Humans: Computational Modeling of Face Recognition. *Network: Computation in Neural Systems* (in press)

Conference Papers

Kleiner, M., C. Wallraven, M. Breidt, D.W. Cunningham and H.H. Bülthoff: Multi-viewpoint video capture for facial perception research. *Captech 2004 - Workshop on modelling and motion capture techniques for virtual environments*, 55-60. (Eds.) Thalmann, N. M. and D. Thalmann (December 2004)

Caputo, B. and C. Wallraven: Object Categorization via Local Kernels. *Proceedings of ICPR 2004*, 132-135 (2004)

Wallraven, C., A. Schwaninger and H.H. Bülthoff: Learning from humans: computational modeling of face recognition. *ECVW 2004* (in press) (2004)

Wallraven, C., B. Caputo and A.B.A. Graf: Recognition with Local Features: the Kernel Recipe. *ICCV 2003 Proceedings 2*, 257-264, IEEE Press (2003)

Bülthoff, H.H., C. Wallraven and A.B.A. Graf: View-based dynamic object recognition based on human perception. *ICPR 2002*, 768-776, IEEE CS Press (2002)

Wallraven, C., Schwaninger, A., Schumacher, S. and Bülthoff, H.H.: View-based recognition of faces in man and machine: Re-visiting Inter-Extra-Ortho. *Biologically Motivated Computer Vision*

2525, 651-660. (Eds.) Bülthoff, H. H., S.W. Lee, T. Poggio and C. Wallraven, Springer (2002)
Wallraven, C. and H.H. Bülthoff: View-based recognition under illumination changes using local features. CVPR 2001 - Workshop on Identifying Objects Across Variations in Lighting: Psychophysics and Computation Proceedings-CD (2001)
Wallraven, C. and H.H. Bülthoff: Automatic acquisition of exemplar-based representations for recognition from image sequences. CVPR 2001 - Workshop on Models vs. Exemplars, IEEE CS Press (2001)
Wallraven, C. and H.H. Bülthoff: Acquiring Robust Representations for Recognition from Image Sequences. DAGM-Symposium München 2001, 216-222, Springer, Berlin (2001)
Wallraven, C., V. Blanz and T. Vetter: 3D-reconstruction of faces: Combining stereo with class-based knowledge. DAGM 1999, 405-412, Springer (1999)

Book Chapters

Wallraven, C. and H. H. Bülthoff: Object Recognition in Man and Machine. Springer, Tokyo (in press)

MPI-Technical Reports

Kleiner, M., C. Wallraven and H.H. Bülthoff: The MPI VideoLab - A system for high quality synchronous recording of video and audio from multiple viewpoints. (123) (May 2004)
Graf, A.B.A. and C. Wallraven: Multi-class SVMs for Image Classification using Feature Tracking. (99) (August 2002)

Abstracts

T. Cooke and C. Wallraven: Implementation of motion detectors: A case study. Interdisziplinäres Kolleg (IK 2003) (March 2003)
Huber, S.A., F.N. Newell and C. Wallraven: Categorisation of dynamic objects. ECVP 2001, Perception (2001)
Wallraven, C., F.N. Newell and S.A. Huber: The Role of Dynamic Object Properties in Categorisation. ESCOP 2001 (2001)

Diploma Thesis

Wallraven, C.: Ein modellbasiertes Stereosystem zur dreidimensionalen Rekonstruktion von Gesichtern. Tübingen (1999)