

A View-Based Approach to Three-Dimensional Object Perception

Dissertation zur Erlangung des Grades
einer Doktorin der Naturwissenschaften
der Technischen Fakultät
der Universität Bielefeld

vorgelegt von

Gabriele Peters

aus Bochum

Dezember 2001

1. Gutachter: Prof. Dr. Christoph von der Malsburg
2. Gutachter: Prof. Dr. Helge Ritter

Tag der Disputation: 13. Februar 2002

Acknowledgements

This thesis was developed at the Institut für Neuroinformatik at the Ruhr-Universität Bochum. I would like to thank all those who contributed to its completion. First of all, Prof. Dr. Christoph von der Malsburg, who guided my research by his criticism and enthusiasm. Furthermore I thank the Ph.D.-committee at the Technische Fakultät of the Universität Bielefeld for their interest in this thesis, especially Prof. Dr. Helge Ritter, who made this cooperation possible.

It was a great time for me working in Prof. von der Malsburg's group and I want to thank all colleagues at the Institut für Neuroinformatik for the extraordinary good atmosphere and for the friendship I have experienced. The cooperative spirit of the group undoubtedly influenced the development of this thesis. I am especially grateful to Dr. Rolf Würtz, who has always been a good advisor and who read the whole manuscript and gave many useful comments. Ingo Wundrich and Dr. Jan Wieghardt read parts of this work and further improved it by their critical remarks. I thank Michael Neef for providing invaluable support by running an almost trouble-free computer network and Uta Schwalm, who was always helpful with administrative problems. My warmest thanks go to Pervez Mirza, who was never tired answering my questions concerning the English language.

In the end, the completion of this work would not have been possible without the support of my parents and my friend Dr. Torsten Wolf.

Contents

1	Introduction	1
2	Theories of Three-Dimensional Object Perception	5
2.1	Theories	5
2.1.1	Volume-Based Representations	5
2.1.2	View-Based Representations	6
2.1.3	Canonical Views	8
2.1.4	Recognition of Unfamiliar Views	9
2.2	Behavioral And Physiological Evidence	9
2.2.1	Evidence for Volume-Based Representations	9
2.2.2	Evidence for View-Based Representations	10
2.2.3	Evidence for Canonical Views	11
2.2.4	Evidence for View Interpolation	12
2.3	Summary And Conclusions	12
3	Preprocessing and Fundamental Techniques	15
3.1	Choice of Objects	16
3.2	Image Acquisition	16
3.3	Segmentation	17
3.4	Gabor Wavelet Transform	19
3.4.1	Gabor Wavelets	20
3.4.2	Gabor Transform	21
3.4.3	Similarity Functions	22
3.5	Labeled Grid Graphs	22
3.6	Matching Local Object Features	23
3.7	Tracking Local Object Features	24
4	Robustness of Views Against Pose Variation	27
4.1	View Bubbles - A Measure of Pose Robustness	27
4.2	Methods of Comparing Matching With Tracking	29
4.2.1	Quantitative Comparison	29
4.2.2	Qualitative Comparison	30
4.3	Results	30
4.3.1	Quantitative Comparison	30
4.3.2	Qualitative Comparison	33
4.3.3	Canonical Views	37

4.4	Discussion	37
4.5	Parallels to Primate Object Perception	39
5	Sparse Object Representation	41
5.1	Generation of a Sparse Object Representation	41
5.1.1	Set Cover Algorithm	42
5.2	Results	43
5.3	Discussion	45
6	Morphed Views	49
6.1	Morphing of Unfamiliar Views	49
6.1.1	Linear Combination of Object Point Positions	50
	Two Sample Views	52
	Three Sample Views	53
6.1.2	Warping From Familiar to Unfamiliar Views	54
6.2	Evaluation of Morphed Views	55
6.2.1	Relative Errors	56
6.2.2	Methods	57
6.2.3	Results	59
6.2.4	Discussion	59
7	Virtual Views	65
7.1	Virtual View Generation	65
7.1.1	Interpolation of Object Point Features	67
	Two Sample Views	67
	Three Sample Views	67
7.1.2	Virtual View Reconstruction	68
7.2	Evaluation of Virtual Views	68
7.2.1	Methods	69
7.2.2	Results	69
7.2.3	Discussion	70
8	Pose and Sequence Estimation	75
8.1	Single Pose Estimation	75
8.1.1	Methods	75
8.1.2	Results	76
8.1.3	Discussion	76
8.2	Sequence Estimation	80
8.2.1	Methods	80
8.2.2	Results	81
8.2.3	Discussion	82
8.3	Adding Noise	84
8.3.1	Methods	84
8.3.2	Results	85
8.3.3	Discussion	86
9	Summary and Conclusions	93

<i>CONTENTS</i>	iii
A Sequences of Matched and Tracked Features	95
B Linear Combination of Object Point Positions	111
B.1 Two Sample Views, x-Coordinate	112
B.2 Two Sample Views, y-Coordinate	113
B.3 Three Sample Views, x-Coordinate	113
B.4 Three Sample Views, y-Coordinate	113
List of Figures	115
List of Tables	117
Bibliography	119
Previously Published Contents of this Thesis	125

Chapter 1

Introduction

One century ago, the french painter Claude Monet painted his great series “Rouen Cathedral”. It consists of more than thirty paintings of the Cathedral at Rouen in the morning, in full sunlight, during the night, at dawn, in the rain, in fog, from the front, and from side viewpoints, and with different details like the portal or the steeple. Although these paintings vary in color, illumination, scale, and viewpoint, and although in some paintings parts of the cathedral are occluded, the beholder is able to recognize the cathedral in each of them (see figure 1.1).

This example illustrates the fundamental mystery of visual perception. Each object in our environment can cause considerably different patterns of excitation in our retinae depending on the illumination or the observed viewpoint of the object. Despite this we are able to perceive that the changing signals are produced by the same object. It is a function of our brain to provide this constant recognition from such inconstant input signals by establishing an internal representation of the object. The nature of such an internal representation and the way how it can be acquired is the concern of scientists of such different disciplines as biology, psychology, physics, mathematics, engineering, and computer science. Since until today no artificial vision system exists that can compete with even simple living systems this is still a highly relevant topic.

This thesis concentrates on a partial problem of object perception, the acquisition and application of a *viewpoint-invariant* object representation. That is a representation which allows the recognition of a three-dimensional object independently from the viewpoint, which is displayed to the observer. There is a present argument about how the brain is able to learn a three-dimensional notion of the environment although the source of information consists of the views projected to the retina, which are only two-dimensional. Two major theories about the nature of object representations are being discussed. On the one hand, some researchers believe that a perceiving system has to establish an internal, three-dimensional model of the object. On the other hand, many scientists are confident that

An object representation in the form of a collection of a few, distinguished views which are connected is sufficient to perform perception functions such as the recognition of the object and the estimation of its pose.

This thesis provides a contribution to the debate which supports the latter model and the

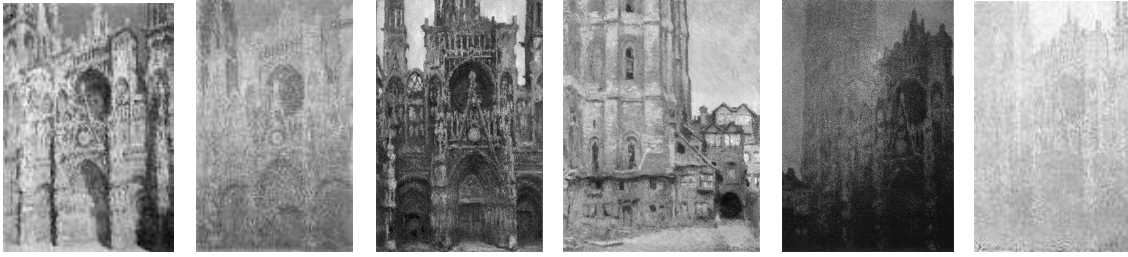


Figure 1.1: Paintings from Monet’s Series “Rouen Cathedral”. These paintings show intense changes in the appearance of the cathedral due to variations in color, illumination, scale and viewpoint. Thus, each painting generates a different pattern of excitation in the eyes of the observer. But our brain is able to generalize from these inconsistencies and to provide a reliable recognition of the cathedral.

cursive phrase can be regarded as the main thesis of this work. In detail, the following questions are dealt with.

- Q1** Can human object perception be appropriately described by a two-dimensional, view-based approach, which copes without an explicit three-dimensional model of the object?
- Q2** If the experience of single object views is sufficient to recognize an object from any other viewpoint, are there distinguished views which may be better suitable for perception than others? Maybe because they are more robust against slight variations in their pose? Are there differences at all between views with respect to their *pose robustness*?
- Q3** If the view-based approach proves to be reasonable, how many views are sufficient to represent a whole object? How are these views distributed on the viewing sphere of the object? How large is the area of *generalization* within which an inference from a sample view to an unfamiliar view is possible?
- Q4** If a collection of a huge number of *unconnected* object views does not seem to be an appropriate form of object representation, what are the strategies to *combine* familiar views?

Chapter 2 (Theories of Three-Dimensional Object Perception) gives an overview of the latest developments in the engineering and computer science disciplines as well as related work in behavioral and physiological research concerning central aspects of this thesis. In the summary of chapter 2 the questions **Q1** to **Q4** are gone through and answers from the related disciplines are summarized. Related work concerning minor aspects of this thesis is covered in the related chapters.

In a nutshell, the aim of this work is the development of an artificial, perceiving system from which answers to these questions can be proposed. The system should be able to independently learn *view-based* (first condition) and *sparse* (second condition) object representations from sample views of real-world objects. Representations of unfamiliar



Figure 1.2: Non-Valid Objects. The left and middle images display examples of degenerate objects this thesis does *not* refer to. On the one hand, the appearance of the sphere is the same for each viewpoint, which makes an estimation of its pose impossible. On the other hand, the core of a digital sundial is an example of an object which rapidly changes its appearance depending on the viewpoint. It projects different light patterns onto a screen depending on the angular position of the sun, which results in a digital display of the time. For this kind of object it is difficult to generalize from sample views to unfamiliar views. The digital sundial's principle of operation is exemplified in the picture taken from the book “Gödel, Escher, Bach” by Hofstadter [23] shown on the right.

views should be derived from those of stored views and the system should be capable of performing perception tasks such as *estimating object poses* (third condition).

To analyze the influence of the viewpoint of an object on the performance of the artificial system it is crucial to use objects which are not degenerate with respect to their appearance depending on the viewpoint. For instance, it would be counterproductive to use objects with equal appearance from many viewpoints. A homogeneously textured sphere is one extreme example of this case. Pose estimation, for instance, is not possible for such objects. On the other hand, an object the views of which change rapidly with the viewing angle would be as intractable. An extreme example of this case is a digital sundial. For these objects a generalization from sample views to unfamiliar views would be difficult (see figure 1.2). Thus, this thesis applies to non-degenerate, but in other respects arbitrary, real-world objects as displayed in figure 1.3.

Chapter 3 (Preprocessing and Fundamental Techniques) describes the acquisition of a densely sampled set of views of such objects and the preprocessing of these images. They provide the data basis for the experiments. Here also some fundamental techniques and concepts are introduced which are needed in later chapters, such as the Gabor wavelet transform, graph matching, and the tracking of object features.

The idea behind representing a three-dimensional object by only a few two-dimensional views is that the chosen views are representative for a preferably large area of surrounding, non-chosen viewpoints. That raises the question of how these areas of pose robustness can be determined. This and other questions from **Q2** are treated in **chapter 4** (Robustness of Views Against Pose Variation), in which two methods to calculate robustness areas of object views as well as correspondences between familiar views are compared. These are essential for the later calculation of unfamiliar views.

Having found a method to determine a surrounding area for each view of an object, where only slight changes in the object's appearance occur, it is easy to derive a sparse, view-based representation by a selection of views the areas of which cover the whole



Figure 1.3: Valid Objects. These are examples of images of real-world objects, i.e. images taken from real objects rather than CAD-generated virtual, ones. This thesis refers to this kind of objects. In fact, these are the objects I have used for my experiments.

viewing sphere. The details of this procedure are described in **chapter 5** (Sparse Object Representation), which contributes to the questions **Q3**.

Question **Q4** is covered by chapters 6 and 7. In **chapter 6** (Morphed Views) unfamiliar views of an object are generated from the sparse representation by a linear combination of object point positions of stored views and a subsequent morphing of image intensities to the calculated, new point positions. As this thesis does not lie in the domain of computer graphics, view morphing only serves as an auxiliary means and technique of visualization to prove the applicability of linear combinations of point positions and to evaluate the sparseness of the object representation.

For the purpose of recognizing an object from an unfamiliar viewpoint the combination of object point *positions* of familiar views is not sufficient. In addition to this, the *features* which describe the local properties of the object point have to be combined as well. This is done by an interpolation method specified in **chapter 7** (Virtual Views).

The ability to calculate the representation of an unfamiliar view (by a linear combination of object point positions and an interpolation of object point features) provides the possibility to estimate the pose of the object, which is displayed from an arbitrary viewpoint, even if the presented test view is degraded by the addition of noise. This procedure is described in **chapter 8** (Pose and Sequence Estimation). The results yielded by the pose estimation also serve to evaluate the quality of the object representation and combination of views.

Finally, in **chapter 9** (Summary and Conclusions) I summarize my results and go again through the questions **Q1** to **Q4**, this time proposing answers which can be derived from my experiments.

Chapter 2

Theories of Three-Dimensional Object Perception

In this chapter current theories of the visual perception of three-dimensional form are introduced (section 2.1). Their plausibility and limitations are discussed with respect to results from behavioral and physiological research which are reviewed in section 2.2.

2.1 Theories

Most of the theories introduced in this section were derived from computer simulations, artificial systems, or technical applications which represent work related to this thesis. The process of acquiring object representations is addressed as well as object recognition and pose estimation. The differences between *volume-* and *view-based* representations¹ are explained and the concepts of *canonical views*, *aspect graphs*, *interpolation*, and *linear combination of views* are introduced.

2.1.1 Volume-Based Representations

Over a long period of time one assumed in the field of cognitive sciences that an explicit three-dimensional model is necessary to recognize an object. The authors argued that most of the objects in the visual world can be divided into one or more volumetric parts, thus it should be possible to represent them by these constituent parts and their spatial relations. Representations based on this principle are called *volume-based* or *model-based*. For example, Nevatia and Binford [42, 43] and Marr and Nishihara [35] were one of the first to propose recognition by reconstruction. According to their model the visual input is totally reconstructed and matched to a three-dimensional representation in memory.

At an earlier period of time Shepard and Metzler [67] already proposed their theory of *mental rotation*. They designed a task in which subjects were shown two novel visual stimuli (random block shapes which were rotated in depth). Their subjects were asked to determine whether the stimuli had the same shape or different shapes. Shepard and

¹The terms *volume-based* and *view-based* representations are mostly used synonymously to the expressions *object-centered* and *viewer-centered* representations, as well as synonymously to *3D-* and *2D-*representations, respectively. A detailed description of these terms can be found in Peters [51].

Metzler argued that subjects, to make their judgment, mentally rotated the shapes in their head until the two stimuli were oriented the same way. Volume-based models mostly explain the generalization from familiar to unfamiliar views by *recognition by alignment* of three-dimensional models, which is strongly connected to the notion of mental rotation. It was proposed by Ullman [73]. During recognition by alignment each stored model undergoes an aligning transformation after which it is compared to the input image. A visual system, which utilizes aligned three-dimensional models, should recognize perfectly, as long as all features used for the transformation are visible.

From a technical point of view, the rotation of a three-dimensional model in a computer (followed by a matching routine) would be a very efficient method for object recognition. However, this requires the availability of a model of the object. It turned out that the acquisition of such a three-dimensional model is either a difficult task which often requires interaction from a user (as, for example, described by Francois and Medioni [19]) or utilizes techniques which have nothing in common with current neuroscience theories. For example, Nevatia and Binford [43] used a laser ranging technique to acquire three-dimensional positions of points on the visible surface of an object.

As for this work biological plausibility is crucial volume-based models of three-dimensional object perception are of minor importance for this thesis.

2.1.2 View-Based Representations

Besides volume-based representations, which seem to be very economical but require the ability of the visual system to transform across views, many computational models have been proposed in which two-dimensional views are combined into the equivalent of a three-dimensional object representation.

The simplest of these view-based descriptions of an object is a densely sampled collection of views which are treated independently. The addition of new views would not increase the complexity of the description, but only increase the size of the search space. Even for such a simple view-based representation the visual system would need the ability to transform views to different viewing angles inside narrow ranges, otherwise an infinite number of views would have to be stored. Simple representations in the form of a collection of *independent* views are unlikely to be realized in the human brain, because otherwise it would be difficult to explain humans' ability to recognize novel views of familiar objects.

More plausible seems to be a representation of an object in the form of a (smaller) collection of *relevant* views only and the spatial relations among them. The relations would preserve the spatial information, which is lost in simpler view-based approaches. Recognition of intermediate views with such a representation could be achieved, e.g., by an interpolation or a linear combination of stored views (see subsection 2.1.4).

A description of three-dimensional objects by *aspect graphs*, proposed by Koenderink and van Doorn [27, 28], is one example of such an advanced view-based representation. The vertices of an aspect graph are constituted by views, which can be interpreted as special points on a transparent viewing sphere with the object in its center. These stored views represent distinct *aspects*, for which an "observer may execute any small movement without affecting the aspect". Relations between aspects are expressed by *events*. Events occur, whenever changes in the viewpoint lead to qualitative changes in the appearance of the object. Here "the aspect changes suddenly if small movements are made" (see

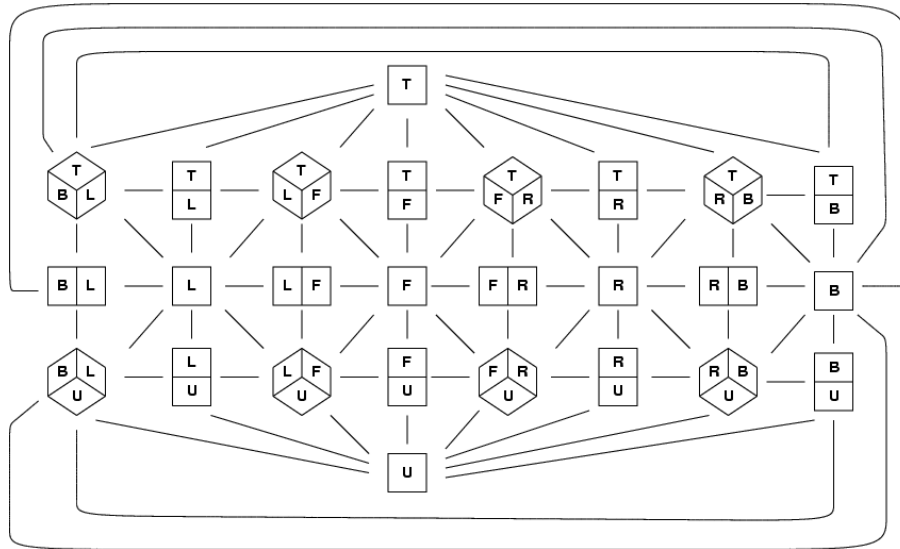


Figure 2.1: Aspect Graph. This illustration of an aspect graph of a cube is taken from Khoh and Kovesi [26]. The sides of the cube are tagged with letters for “left”, “right”, “top”, and so on. Each vertex of the graph represents a single aspect of the cube. The graph edges represent the events, which mark the possible transitions between aspects.

figure 2.1).

Aspect graphs have been applied, e.g., by Seibert and Waxman. In [62] they describe the learning of representations for objects from arbitrary sequences of the rotating object. Based on edge and corner detection, they cluster the views into different aspects. Each object is then represented by a “transition matrix”, which contains the probability for the transition from one aspect to another. Utilizing these transition matrices their system is able to recognize objects, but pose estimation or the generation of intermediate (non-experienced) views is not possible.

Another view-based approach, which does allow pose estimation, was proposed by Murase and Nayar [41]. They represent objects by a manifold in eigenspace. An input image of an object to be recognized is projected onto the eigenspace of the learned objects. The object is recognized based on the manifold it lies on. The exact position of the projection on the manifold determines the object’s pose. Manifolds became a very popular idea in recent years for various kinds of applications which require dimensionality reduction of large data sets. For instance, Roweis and Saul [60] applied it to arrange words extracted from encyclopedia articles in a continuous semantic space. Especially for view-based object recognition the manifold approach seems to be a promising method as, e.g., Tenenbaum et al. [70] and Seung and Lee [66] suggest. A system which belongs to this category is proposed by Wiegardt [81], who has shown that an object representation can be derived from single views. The learned representation preserves the topology of the views and allows a coherent description of the object’s appearance from unfamiliar view points.

Besides these algebraic methods general purpose learning methods have been applied to view-based object recognition. Support vector machines, for example, are successful in recognizing three-dimensional objects from two-dimensional views (Blanz et al. [6],

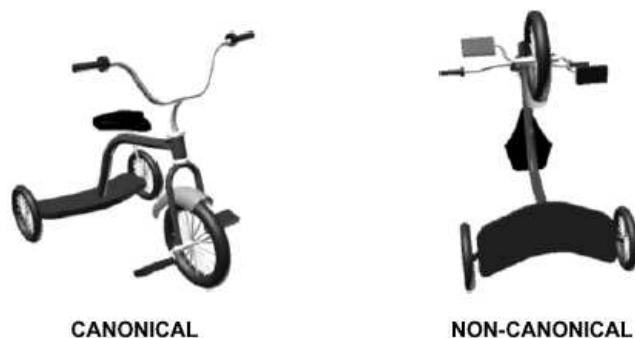


Figure 2.2: Canonical Views. This example of a canonical and a non-canonical view of a tricycle is taken from Bülthoff et al. [10].

Schölkopf [61]). Further view-based approaches, e.g., by Poggio and Edelman [56] and Ullman and Basri [74], are described in subsection 2.1.4.

2.1.3 Canonical Views

Across the different models for three-dimensional object perception, the notion of a *canonical view* is a prominent topic. It can be regarded as a view which is easier to recognize than other views of the same object. A hard definition does not exist, even its properties are controversial. Palmer et al. [45] describe canonical views as the ones that “humans find easiest to recognize and regard as most typical”. Often the term *characteristic view* is used synonymously (see figure 2.2). Many explanations have been provided as to why some views are easy to recognize and others are not. Different features have been considered for this classification. One aspect for defining canonical views might be the constellation of visible corners and edges. For example, Gray [20] clustered the viewing sphere of line drawings of geometrically faceted objects into regions of similar views based on corners and edges. He obtained nine clusters, i.e., nine canonical views, which represent a whole object.

Open questions concerning canonical views are the number of views necessary for different visual tasks and their statistical distribution on the viewing sphere. Malik and Whangbo [34], for instance, have demonstrated that an uniform distribution is inappropriate. Weinshall and Werman [77] have shown that the likelihood to observe a certain view of an object correlates with the view’s robustness against pose variation, i.e., how little the image changes when the viewpoint is slightly changed. The most likely views are often the “flattest” views of an object.

Although canonical views are more strongly connected with view-based models volume-based models also have to explain the phenomenon of canonical views. One possibility is the concept of *salient* or *non-accidental* features used, e.g., by Lowe [33] or Biederman [5]. In this scheme some parts of an object are particularly salient and their visibility facilitates recognition. Accordingly, canonical views are not derived from a general procedure, rather they highly depend on the specific object.

2.1.4 Recognition of Unfamiliar Views

Starting from the assumption that canonical views play a dominant role in object recognition the question arises how non-canonical views and views which have not been experienced before and are not stored in the representation can be recognized.

One widespread approach to the generalization from familiar to unfamiliar views is the *interpolation* of unfamiliar views. Novel views can be generalized from stored views by view approximation as described, e.g., by Poggio and Edelman [56]. According to this theory, humans and other primates can achieve viewpoint-invariant recognition of objects by a system that interpolates between a small number of stored sample views. This model predicts that unfamiliar views lying between stored views are recognized easier than those which are somewhere else on the viewing sphere. In addition, recognition rates should deteriorate with an increasing distance of the novel view from a stored view. Seitz and Dyer [63] have shown that under certain assumptions about visibility image interpolation is a physically valid mechanism for view interpolation, which means that the interpolation between two views of an object produces a physically valid intermediate view of it.

Another theory about the recognition of unfamiliar views, which also belongs to the view-based approaches, is the recognition by a *linear combination* of views. Ullman and Basri [74] showed mathematically that, under orthographic projection, the two-dimensional coordinates of an object point for a special view can be expressed as a linear combination of its coordinates in a limited set of other viewpoints, provided that the correspondences between points in all views are known and no self-occlusions occur. The number of required views depends on the complexity of the object and the allowed three-dimensional transformations. In contrast to the interpolation model, this model predicts equally high recognition rates for unfamiliar views lying in the space spanned by the stored views, independent of the distance between novel and stored views. Among others, Beymer and Poggio [4] used a linear combination approach to apply prior knowledge of an object class (faces) to generate virtual views for face recognition.

2.2 Behavioral And Physiological Evidence

In this section results of studies with humans and monkeys are described which either support or contradict the theories of three-dimensional object recognition introduced in the last section.

2.2.1 Evidence for Volume-Based Representations

Behavioral and physiological evidence for a volume-based coding of objects is sparse. There are nevertheless hints from physiological studies which contradict a purely view-based approach. In the temporal cortex cells have been detected which exhibit object-centered coding, i.e., they respond equally to a large number of views of an object. They were detected in macaque brains, e.g., by Perrett et al. [48] for the coding of heads and by Booth and Rolls [8] for the coding of small plastic objects. In [48] they found cells selective to a large number of views of one individual's head, but unresponsive to all tested views of a different individual. In [8] neurons are described which were responsive to all views of an object, providing evidence that these neurons were coding for objects,

rather than for individual views. However, this is not necessarily an argument for a volume-based coding of objects. These and other results reported in [48] and [8] as well as results reported in the following subsections only suggest that three-dimensional object recognition in the primate brain operates in a hierarchical fashion with increasing levels of abstraction. Starting with size- and viewpoint-specific representations, to size-invariant, but viewpoint-specific codings, up to viewpoint-invariant representations.

2.2.2 Evidence for View-Based Representations

There are uncountable behavioral studies with primates that support the model of a view-based description of three-dimensional objects by our visual system. For instance, if a set of unfamiliar object views is presented to humans their response time and error rates during recognition increase with increasing angular distance between the learned (i.e., stored) and the unfamiliar view, as Edelman and Bühlhoff [16] point out. (This decrease in recognition speed for unfamiliar views was already reported by Tarr and Pinker [69] and quite early by Bartram [1]. For monkeys the ability to generalize from training views was found by Logothetis et al. [30, 31] to deteriorate with increasing rotation angle.) However, this angle effect on performance declines if intermediate views are experienced and stored, as Tarr [68] reports. In addition, Cutzu and Edelman [14] analyzed the response time and error rate scores and found out that they are not linearly dependent on the shortest angular distance in three dimensions to the best-recognized view, as predicted by the mental rotation theory. Rather, the performance was correlated with an “image-plane feature-by-feature deformation distance” between the test view and the best view. This leads them to the suggestion that the measurement of image-plane similarity to a few feature patterns is an appropriate model for human three-dimensional object recognition.

Experiments with monkeys were made by Logothetis et al. [31] which showed that familiarization with a “limited number” of views of a novel object can provide viewpoint-independent recognition. Three views of a wire-like object, 120° apart, often were sufficient for recognizing any view resulting from rotations around the same axis. For the entire viewing sphere about 10 views were sufficient to achieve view-independent performance. But the same study claims that the number of required views may depend on the object class. It may reach a minimum for a novel object of a familiar class, e.g., for a new individual face one view only may be sufficient. The inability of monkeys to recognize objects rotated by more than approximately 40° from a single familiar view is also reported.

A similar result for human object recognition was published earlier by Rock and DiVita [59]. Subjects performed very poorly in recognizing wire-like objects for view distances larger than approximately 30° . They could not even imagine how the objects would look when rotated further.

In psychophysical studies Kellman [25] found out that adult humans are able to perceive three-dimensional form from static views of objects. He suggests that this ability leans on extrapolations to the whole form based on simplicity or symmetry considerations, which may be products of learning. In [80] Wexler et al. report a psychophysical experiment in which subjects were instructed to perform mental rotation, but they switched spontaneously to landmark-based strategies, which turned out to be more efficient.

Numerous physiological studies also give evidence for the existence of a view-based processing of the brain during object recognition. Some of them are listed in subsection 2.2.3

about canonical views. Logothetis et al. [32], for example, made recordings of single neurons in the inferior temporal cortex (IT) of monkeys, which is known to be concerned with object recognition. The results of these recordings resemble those obtained by the behavioral studies mentioned above. They found populations of IT neurons which responded selectively to only some views of a previously unfamiliar object and their response declined gradually as the object was rotated away from the preferred view.

2.2.3 Evidence for Canonical Views

The controversial properties of canonical views are reflected by two contradictory studies. On the one hand, in a study by Palmer et al. [45] subjects had to choose canonical views for several objects. In the views they selected, the object's principal axis often was 45° to the line of sight. On the other hand, Perrett and Harries [46] found out that humans prefer views with the principal axis of the object either parallel or perpendicular to the line of sight. These views can be referred to as "side" and "end" views, respectively, which are often "plan views", equivalent to those drawn by an architect to represent an object. (This result is consistent with Weinshall and Werman's "flattest" views mentioned in subsection 2.1.3.) The benefit of such views is the absence of perspective distortions in the third dimension, as they point out.

Nevertheless, there is no doubt that most real objects possess views which are easier to recognize than others. This was confirmed early by a report of Warrington and Taylor [76] who observed patients with lesions in the right parietal cortex. These patients performed worse than control subjects in recognizing objects from "unusual" views, whereas "usual" views were not affected. This indicates a different processing of canonical and non-canonical views in the brain.

Many experiments were carried out by Bülthoff et al. [16, 10]. They confirmed that naming was fastest if a stimulus was in a canonical view. These views were established even if in the training phase each view of an unfamiliar object appeared with equal frequency. In later experiments carried out by Blanz et al. [7] participants had to mentally imagine an object on the one hand and, on the other hand, had to adjust it to the viewpoint from which they would take the "best" photograph to illustrate a brochure. Both tasks yielded almost the same views and there was a large degree of consistency across the participants.

Physiological studies also provide evidence for the existence of canonical views. Cells in the temporal cortex of macaques have been found which respond selectively to faces, hands, and other classes of biologically significant objects. The majority of these cells exhibits a view-based response pattern, i.e., some of them respond selectively to face or profile views of heads, as described by Perrett et al. [50], although at the same time they generalize across image position, size, orientation in the image plane, color, and lighting conditions (Hietanen et al. [22]). There are more cells optimally tuned to canonical views like full face or profile views than to other views (Perrett et al. [48]). The tuning covers views between 45° and 70° from the cells' optimal views until the response is reduced to half of its maximum. Interestingly, the same views seem to be important physiologically and behaviorally. Also the relative importance of views is comparable. Face and profile views appear more important than half-profile views, and all of these front views are more important than rear views of a head, both in behavioral and physiological studies, as reported by Perrett et al. [49].

2.2.4 Evidence for View Interpolation

Bülthoff and Edelman [9] made psychophysical experiments to compare view interpolation, linear combination of views, and alignment of three-dimensional models for the recognition of unfamiliar views. Subjects were shown two training views of a computer-generated three-dimensional wire-like object, which were 75° apart. In the test phase a novel view was presented, which was either on the same rotation axis *between* the training views, or on the same rotation axis *beyond* them, or on an axis *orthogonal* to the training axis. The error rates during recognition mostly fit the predictions of the interpolation model, i.e., the error rates were lowest for the *between* condition, medium for the *beyond* condition and highest for the *orthogonal* condition. According to the authors, this contradicts a pure linear combination model, which predicts the same good performance for the *between* and *beyond* condition and poor performance for the *orthogonal* condition. However, the disavowal of the theory of a linear combination of views can be maintained only subject to the assumptions of perfect correspondences between sample views and the absence of self-occlusions, which are not necessarily fulfilled under realistic conditions.

The experiment carried out by Bülthoff and Edelman also contradicts alignment models, which predict uniformly good performance for all three test conditions. Anyway human perception of Euclidean metric structures is very limited, a general observation which does not support the idea of precise recognition by alignment. For instance, Todd and Reichel [72] have shown that humans perform poorly on estimating the precise distance between two points in the environment, but are nevertheless able to determine which point is closer to them.

Results similar to the ones reported by Bülthoff and Edelman were obtained in physiological studies by Logothetis et al. [30]. They trained monkeys with two views of a computer-rendered wire or spheroidal novel object, which were far apart, e.g., 120° . The monkeys were able to recognize all test views inside this interval, whereas the extrapolation along either the same or an orthogonal axis was limited.

2.3 Summary And Conclusions

Many contradictory studies demonstrate the difficulty to put forward a clear statement about the perception of three-dimensional objects. Probably the brain of humans and other primates is able to perform both, recognition based on object-centered as well as viewer-centered representations. The choice of the appropriate form may depend, e.g., on the structure of the object and the task to be performed. For example, if an object is to be grasped a volume-based representation is more appropriate than a view-based one. However, concerning view-based representations the following answers to the questions posed in the introduction (chapter 1) can be derived from the performances of artificial and living systems.

- Q1** Human object perception and that of other primates can be interpreted by a *view-based* approach. Object representations in form of single, but connected views are sufficient for a huge variety of situations and perception tasks such as object recognition and pose estimation.

- Q2** Real-world objects possess *canonical views*, which are better suitable for recognizing the object than other, non-distinguished, views. The robustness of views against pose variation can be expressed in terms of *aspects*, which are surrounding areas on the viewing sphere without qualitative changes in the object's appearance.
- Q3** The number of views which are sufficient to represent an object depends on the specific object. The representing views are not distributed uniformly on the viewing sphere, rather their distribution depends on the object. Primates are able to generalize from a reference view to surrounding, unfamiliar views. A lower bound of about 30° and an upper bound of about 70° distance from the reference view can be found in the literature for this *generalization*, but a distance of about 30° to 40° seems to be appropriate.
- Q4** Presumably a kind of *view interpolation* occurs between familiar views when unfamiliar views have to be recognized. The disavowal of the theory of a *linear combination of views* by psychophysical studies derives from experiments carried out with wire-like objects. As these objects differ from real-world objects, e.g., with respect to self-occlusions the linear combination approach is still worth to be explored in object perception from images of real objects. Both combining techniques, nevertheless, require *correspondences between sample views*.

Chapter 3

Preprocessing and Fundamental Techniques

In this chapter the choice of test objects used in this thesis and the acquisition and preprocessing of the image data bases are described. In addition, some fundamental techniques are introduced which are needed in the following chapters.

An appropriate choice of test objects is crucial to draw generally valid conclusions from the results of experiments. The objects used throughout this thesis are introduced in section 3.1. For the establishment of a sparse object representation as well as for providing ground truth data for the evaluation of the quality of reconstructed morphed and virtual views and estimated object poses a dense sampling of the viewing hemisphere of an object is necessary.¹ The acquisition of such dense image data bases is described in section 3.2. The image acquisition provides a large number of images of the object. Each image displays a different view of it and parts of the background. Further processing requires the generation of a description for each of the recorded images which represents the view of the object without the background. For this purpose the object has to be separated from the background in each image. This is described in section 3.3. After the object has been isolated in an image a grid graph is put onto the segment which has been assigned to the object (see figure 3.1). At each vertex of the graph Gabor filter responses are extracted which describe the local surroundings of the vertex. The Gabor transform is described in section 3.4 and the generation of grid graphs in section 3.5.

The representation of each view in form of a labeled grid graph provides the basis for two techniques which are used later (matching and tracking of local object features). They are described in the last two sections 3.6 and 3.7. Both techniques are applied to assess the similarity between different object views and to find corresponding object points in different views. With respect to these applications they will be compared later in chapter 4. The tracking procedure will be used to establish the sparse representation of a three-dimensional object, which is described in chapter 5. The matching technique will be applied to the pose estimation of single views and sequences utilizing virtual views, which is described in chapter 8.

¹Throughout this thesis only the upper viewing hemisphere is considered, but the methods and results can be generalized to the whole viewing sphere of an object.

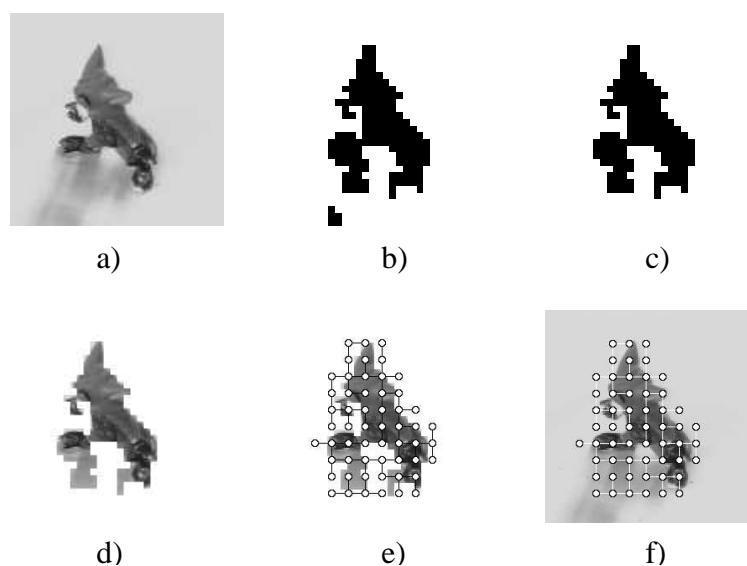


Figure 3.1: Preprocessing. a) Recorded original image, b) Result of the segmentation, c) Result after eliminating wrong segments, d) Original image masked with resulting segment, e) Grid graph covering the object, f) Grid graph shown on the original image.

3.1 Choice of Objects

As the acquisition of object views, which is described in the next section, is a very difficult and time-consuming process, the results of this work have been derived from only two test objects. Thus, the properties of the objects should provide challenging problems to the perception system. The objects used in the experiments of this work have been chosen with this in mind. They are shown in figure 3.2 and differ in the degree of their complexity. The “dwarf” object is a simple object, whereas the “Tom” object is more complex. “Simple” means that the views of the object do not change rapidly while the object rotates. Such objects are often difficult to deal with as already mentioned in chapter 1. The “dwarf” is a relatively convex object with a rather similar shape for all viewing directions, whereas “Tom” is a more irregular object with faster changing views. Both objects vary in the degree of self-occlusions, which occur earlier for the “dwarf”, because of its sphere-like shape. This phenomenon is illustrated in figure 3.2 and will be important in later chapters.

3.2 Image Acquisition

To acquire the views of the upper hemisphere of an object I used an anthropomorphic robot, which has a redundant manipulator arm with seven degrees of freedom, kinematics similar to a human arm and a parallel jaw gripper, described by Becker et al. [2]. The object was fixed on a small piece of squared timber which was placed in the gripper of the robot, while the squared timber and the gripper were covered by gray paper to achieve a homogeneous background (see figure 3.3). I wrote a program which caused the arm to move in a way that the object rotated around the center of its footpoint while a fixed

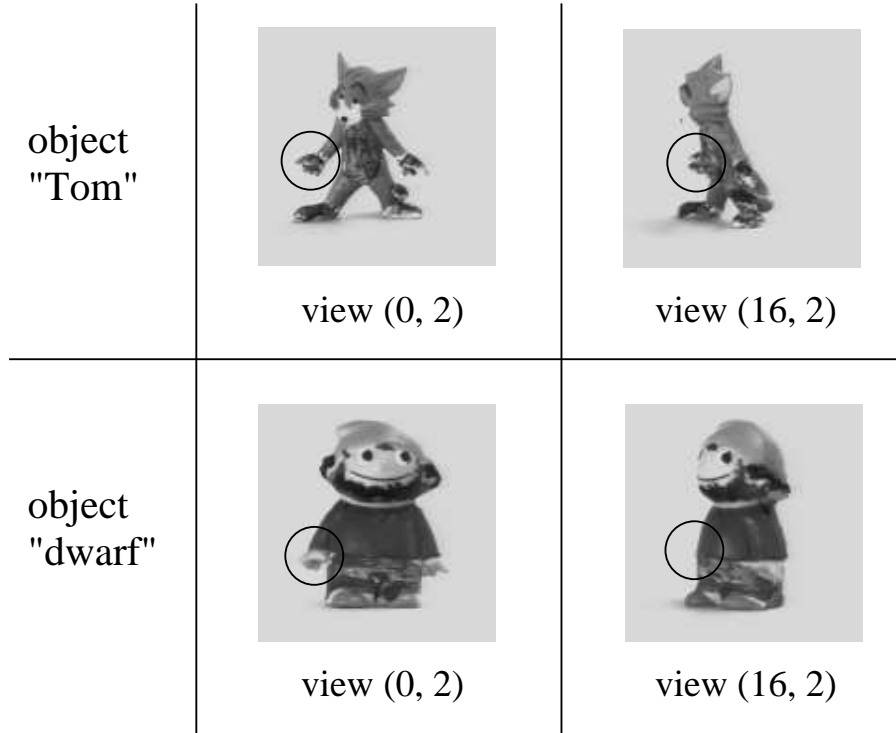


Figure 3.2: Self-Occlusions - Difference Between Object “Tom” and Object “Dwarf”. Self-occlusions of parts of the object occur earlier during rotation for the “dwarf” object than for the “Tom” object, because the “dwarf” is a more compact object. In these examples both objects are rotated with the same angle from the frontal position to the right. The right hand marked by circles has vanished for the “dwarf” object, whereas it is still visible for the “Tom” object.

PULNiX TM-9700 monochrome CCD camera recorded the views as gray level images of size $N \times M = 128 \times 128$ pixels with 256 gray levels. The arm moved in a stepwise fashion to cover the upper viewing hemisphere of an object by 2500 views: 25 equidistant views on each line of longitude and 100 equidistant views on each line of latitude (with a distance of 3.6°). A view is denoted by its position index (p, q) and H denotes the set of all views: $H := \{ (p, q) \mid p = 0, \dots, 99, q = 0, \dots, 24 \}$ (see figure 3.4).

3.3 Segmentation

The segmentation method is based on a system developed by Vorbrüggen [75] and described by Eckes and Vorbrüggen [15]. The segmentation model contains *Potts* spins with coarse-to-fine dynamics comparable to renormalisation methods often used in theoretical physics. Average intensity is used as the only low-level cue, although the system is able to make use of additional cues if they become available.

The segmentation model divides an incoming image of some fixed resolution into P small patches I_i , $i = \{1, 2, \dots, P\}$. Each patch receives a label s_i , $i = \{1, 2, \dots, P\}$ that encodes its membership of one of several possible segments (see figure 3.5 a)–c)). Because of the analogy between this label-based model and an interacting spin system in solid



Figure 3.3: Robot Scene. The robot arm has the squared timber on which the “Tom” object is fixed in its gripper. The timber and the gripper are covered by a gray background. (For the recording of the images not the displayed stereo cameras were used but another single camera.)

state physics, such a label is called a *spin*. The range k of values allowed for a spin $s_i \in \{1, 2, \dots, k\}$, is a parameter of the system and is set to $k = 2$, because only two segments have to be separated, the object and the background. $P = 32 \times 32 = 1024$ patches are used, resulting in patches of $4 \times 4 = 16$ pixels for the images of size 128×128 . The aim now is to find the spin configuration which encodes the “correct” segmentation of the given scene. Each spin s_i interacts with all other spins s_j via an interaction matrix W_{ij} . The difference in mean intensity $|\bar{I}_i - \bar{I}_j|$ at the corresponding image regions is used to compute the interaction W_{ij} between the two spins s_i and s_j assigned to these positions. The desired segmentation is mapped onto the global minimum of the following energy function:

$$E(s) = -\frac{1}{2} \sum_{i=1}^P \sum_{j=1, j \neq i}^P W_{ij} \cdot \delta_{s_i, s_j} \quad \text{with} \quad (3.1)$$

$$W_{ij} = \max\left(1 - \frac{|\bar{I}_i - \bar{I}_j|}{\alpha}, 0\right) - \bar{W}. \quad (3.2)$$

The parameter α ($\alpha = 100$ is used) in combination with the maximum function ensures that the difference in average intensity on the interval $[0, \alpha]$ is mapped to $[0, 1]$. To stress the *Gestalt* law of neighborhood introduced by Wertheimer [79], the interaction is restricted to patches with distances below 7.1 patches. In order to map low similarity to negative interaction and high similarity to positive one, the mean interaction \bar{W} is subtracted from all similarity values which provides the used interaction matrix W_{ij} . The *Metropolis* algorithm developed by Metropolis et al. [38] is used at zero temperature with coarse-to-fine dynamics to let the system relax to a local energy minimum (see figure 3.5 d) and [15] for details). Three stages and $P(1) = 1024$ have been used as the number of

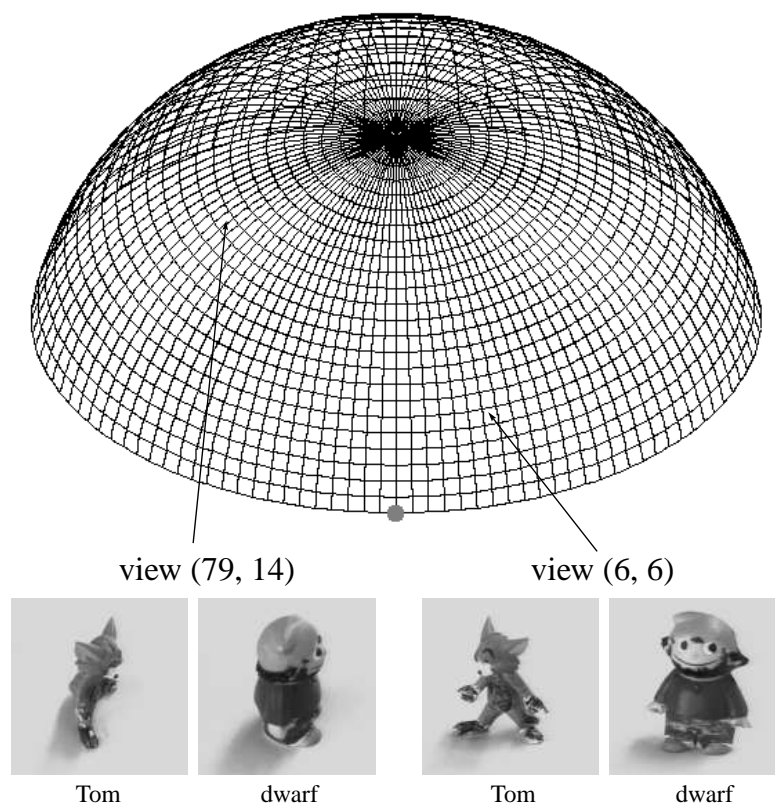


Figure 3.4: Viewing Hemisphere. The image data base for each object consists of 100×25 views which cover the upper viewing hemisphere. Each crossing of the grid stands for one view. The dot in front marks view $(p, q) = (0, 0)$. Two views of two sample objects are shown.

patches in the highest resolution. The number of patches in each resolution is given by $P(n) = P(1)/4^{(n-1)}$.

The segmentation as described may also provide regions, which are erroneously regarded as belonging to the object due to their gray levels like shadows in the background. If the segmentation process yields a non-contiguous object segment the wrong segments are discarded by simply choosing that segment as object which is closest to the center of the image. Figure 3.1 c) shows a typical result of this “centered segmentation”.

3.4 Gabor Wavelet Transform

To derive local descriptions for the recorded images each of the original images undergoes a Gabor wavelet transform, i.e., it is convolved with a family of Gabor kernels.

Gabor wavelets are used in computer vision because they seem to approximate the response patterns of neurons in the visual cortex of mammals as proposed by Jones and Palmer [24]. Burr et al. [11] found out that in the early stages of visual processing neurons display sensitivity profiles in pairs with even and odd symmetry which can be related to the real and imaginary part of a Gabor wavelet.

A formal definition of Gabor wavelets is given in the next subsection, then a description

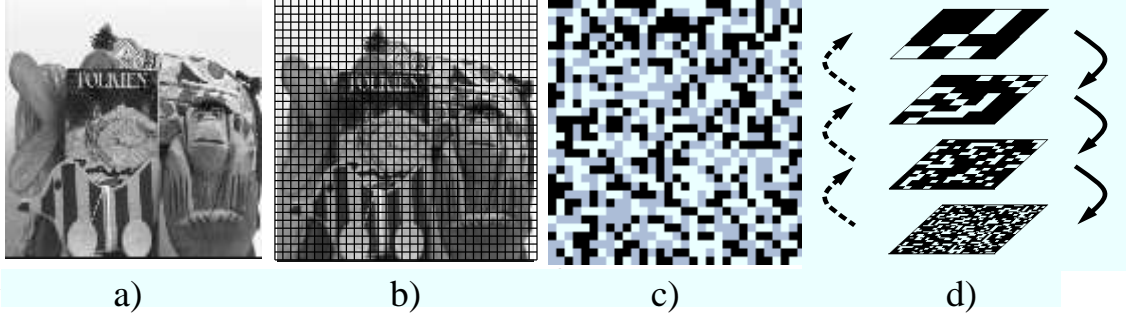


Figure 3.5: Segmentation Model. a) Complex scene to be segmented, b) Scene divided into $32 \times 32 = 1024$ patches, c) Corresponding randomly initialized spin image for $k = 3$, in which each spin value is displayed as the appropriate gray level, d) Renormalisation of the interaction between spins on different resolution levels (arrows on left) and coarse-to-fine dynamics (arrows on right). (adapted from [15])

of the convolution of an image follows (including the introduction of the fundamental *jet* concept), and finally some similarity functions are introduced, which are used later. The technical descriptions in this section are partly taken from Lades et. al. [29].

3.4.1 Gabor Wavelets

The Gabor kernels in image coordinates take the form of a plane wave restricted by a Gaussian envelope function:

$$\psi_{\vec{k}}(\vec{x}) = \frac{\vec{k}^2}{\sigma^2} \exp\left(-\frac{\vec{k}^2 \vec{x}^2}{2\sigma^2}\right) \left[\exp(i\vec{k}\vec{x}) - \exp(-\sigma^2/2) \right]. \quad (3.3)$$

The parameter \vec{k} determines the wavelength and orientation of the kernel $\psi_{\vec{k}}$ and the width of the Gaussian window. The parameter σ determines the ratio of window width to wavelength, i.e., the number of oscillations under the envelope function. The first term in the square brackets determines the oscillatory part of the kernel. The second term compensates for the dc-value of the kernel, to avoid unwanted dependence of the filter response on the absolute intensity of the image. The complex valued $\psi_{\vec{k}}$ combine an even (cosine-type) and odd (sine-type) part (see figure 3.6). The kernels can be sampled with L frequencies and D orientations according to

$$\vec{k}_{\nu\mu} = k_{\nu} \cdot \begin{pmatrix} \cos \phi_{\mu} \\ \sin \phi_{\mu} \end{pmatrix} \quad \text{with} \quad k_{\nu} = \frac{k_{max}}{f^{\nu}}, \quad \phi_{\mu} = \frac{\pi\mu}{D}, \quad (3.4)$$

$\nu \in \{0, \dots, L-1\}$, $\mu \in \{0, \dots, D-1\}$, $k_{max} = \pi/2$ and $f = \sqrt{2}$, which is the spacing factor between kernels in the frequency domain. For the Gabor transform of the recorded images I chose a family of kernels with $L = 4$, $D = 8$, and $\sigma = 2\pi$.

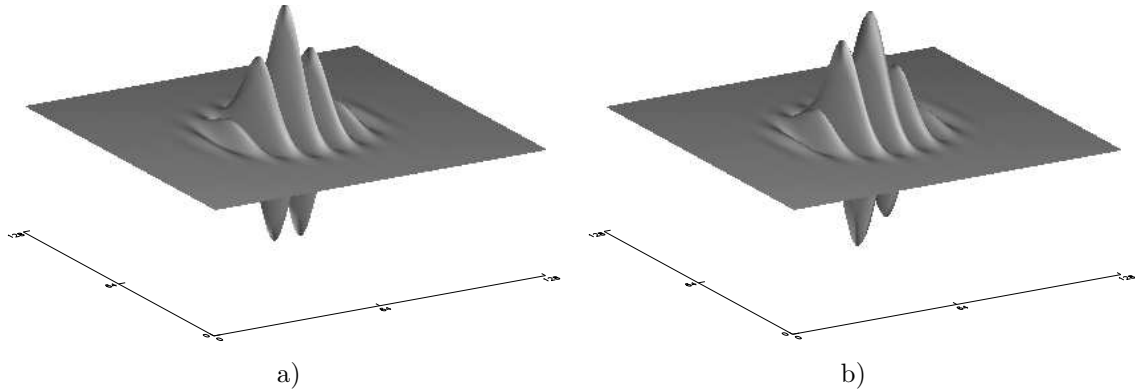


Figure 3.6: Shape of a Gabor Wavelet. a) Real part (cosine phase). b) Imaginary part (sine phase). All kernels have the same shape except for size and orientation. (adapted from [29])

3.4.2 Gabor Transform

If $I(\vec{x})$ is the gray level distribution of the input image, the operator \mathcal{W} symbolizes the convolution:

$$(\mathcal{W}I)_{\vec{k}}(\vec{x}_0) := \int \psi_{\vec{k}}(\vec{x}_0 - \vec{x}) I(\vec{x}) d^2x = (\psi_{\vec{k}} * I)(\vec{x}_0). \quad (3.5)$$

Accordingly, at each image coordinate a filter responses for each of the $L \times D$ Gabor wavelets is obtained.² Filter responses at one image coordinate \vec{x}_0 form a *jet*

$$\mathcal{J}_{\vec{k}}(\vec{x}_0) := (\mathcal{W}I)_{\vec{k}}(\vec{x}_0). \quad (3.6)$$

As the result of the convolution is complex, the i th component of a jet can be expressed in terms of amplitude a_i and phase ϕ_i : $\mathcal{J}_i = (a_i, \phi_i)$ for $i = 1, \dots, L \cdot D$. A jet can be regarded as a descriptor of the local surroundings of the point \vec{x}_0 in the input image (see figure 3.7).

Jets provide the basis for further processing of object views, e.g., they are use for comparing different views of one object (section 3.6), or for tracking object points along changing viewpoints (section 3.7). They are the fundamental data structure on which I base the sparse object representation (chapter 5). Furthermore, virtual views are generated by varying two parameters, the *positions* of object points and the *features* describing the surroundings of these points (chapter 7). These features are jets, too. In addition, the reconstruction of views (subsection 7.1.2) and pose and sequence estimation (section 8) are founded on jets.

In many of these applications it is necessary to determine the similarity between two jets. For this purpose similarity functions have been proposed, which are introduced in the next subsection.

²To reduce processing time the computations are performed in Fourier space, i.e., the kernels are computed in the frequency domain, the input image is transformed via *FFT* (fast fourier transform), then, according to the convolution theorem, only a multiplication of image by kernel is required instead of solving an integral. Finally, the result of the multiplication is backward transformed via *FFT*⁻¹.

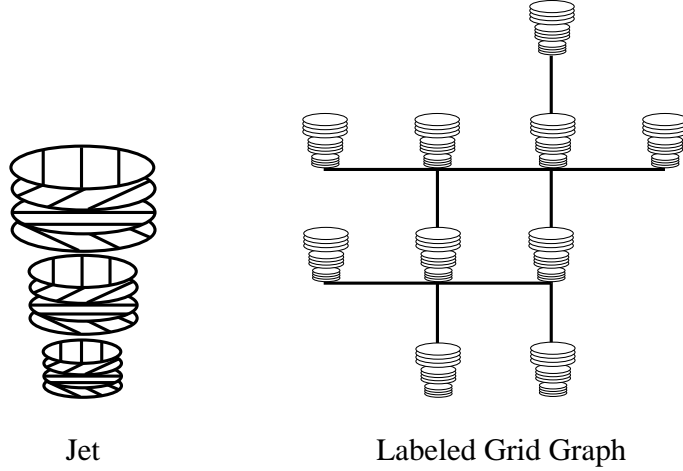


Figure 3.7: Jet and Labeled Grid Graph. This jet is generated from a family of Gabor kernels with 3 frequencies and 4 directions. Each vertex of a graph will be labeled with such a jet. This is described in section 3.5.

3.4.3 Similarity Functions

Two similarity functions between two jets \mathcal{J} and \mathcal{J}' have been suggested by Lades et al. and Wiskott [29, 82]. Whereas \mathcal{S}_{abs} uses the magnitudes of the jets only, \mathcal{S}_{pha} takes the phases of the jets into account as well:

$$\mathcal{S}_{abs}(\mathcal{J}, \mathcal{J}') = \frac{\sum_i a_i a'_i}{\sqrt{\sum_i a_i^2 \sum_i a_i'^2}}, \quad (3.7)$$

$$\mathcal{S}_{pha}(\mathcal{J}, \mathcal{J}') = \frac{1}{2} \cdot \left(\frac{\sum_i a_i a'_i \cos(\phi_i - \phi'_i)}{\sqrt{\sum_i a_i^2 \sum_i a_i'^2}} + 1 \right). \quad (3.8)$$

Both similarity functions have the range $[0.0, 1.0]$. \mathcal{S}_{abs} provides rather similar values for neighbouring jets, if the position of one of the jets is shifted only slightly in the image. In contrast, \mathcal{S}_{pha} provides rapidly changing values in this situation, because the phases of filter responses change significantly even in close neighbourhoods.

3.5 Labeled Grid Graphs

As described earlier each view of the densely sampled viewing hemisphere of an object is to be represented by a labeled grid graph.

In section 3.3 it has been described how an object is separated from the background by a segmentation algorithm. Given the result from the “centered segmentation” I mask the original image with it (see figure 3.1 d)) and cover the masked image with a grid graph with equidistant $X \times Y$ vertices starting at pixel position $(0, 0)$ with a distance of $\lfloor \frac{N-1}{X-1} \rfloor$ pixels in x - and $\lfloor \frac{M-1}{Y-1} \rfloor$ pixels in y -direction. $X \times Y = 13 \times 13$ is chosen.

Then all vertices of the graph are deleted which lie on the background and have a distance to the object segment which is $\geq \lfloor \frac{N}{X} \rfloor$ pixels. The deletion of vertices leads to a

graph \mathcal{G} which covers the object in the image, including its outline, which carries much structural information, while at the same time incorporating only little information on the background (see figure 3.1 e) and f)). Each of the remaining vertices is then labeled with the jet $\mathcal{J}_k(\vec{x})$, which corresponds to the position \vec{x} of the vertex k in the original image, as described in subsection 3.4.2. In other words, each vertex of graph \mathcal{G} is positioned on an object point and equipped with a description of the local surroundings of this point (except a few vertices which lie on the background).

For display purposes the vertices of the resulting graph are connected by a minimal spanning tree (see figure 3.1 e) and f)). These edges are not used for computations. The computationally relevant aspects of graphs throughout this thesis are the *positions* \vec{x}_k of the vertices k in the image and the *features* \mathcal{J}_k attached to the vertices for $k = 1, \dots, n$ if n is the number of vertices of the graph. A graph which represents view (p, q) on the viewing hemisphere is denoted by $\mathcal{G}_{(p,q)}$ for $p = 0, \dots, 99$ and $q = 0, \dots, 24$. It can be expressed in the following way:

$$\mathcal{G}_{(p,q)} = \langle \mathcal{X}_{(p,q)}, \mathcal{F}_{(p,q)} \rangle \quad (3.9)$$

with $\mathcal{X}_{(p,q)} = \{\vec{x}_k\}_{k=1, \dots, n}$ the set of vertex positions and $\mathcal{F}_{(p,q)} = \{\mathcal{J}_k\}_{k=1, \dots, n}$ the set of the corresponding feature vectors.

3.6 Matching Local Object Features

In this section one of the techniques is described which is used to measure the similarity between different object views. *Elastic Graph Matching* has initially been developed for object recognition. It is described in detail in [29]. Given a graph \mathcal{G} with vertices labeled with jets $\mathcal{J}^{\mathcal{G}}$, the aim of matching this graph to an image I is to find new vertex positions which optimize the similarity of the vertex labels to the features extracted at the new positions. During the matching process graph \mathcal{G} is distorted resulting in a graph called \mathcal{G}^I . If its jets extracted from the image are called \mathcal{J}^I the total similarity between \mathcal{G} and \mathcal{G}^I is computed as averaged similarity for each vertex:

$$\mathcal{S}_{tot}(\mathcal{G}, \mathcal{G}^I) = \frac{1}{n} \cdot \sum_{k=1}^n \mathcal{S}(\mathcal{J}_k^{\mathcal{G}}, \mathcal{J}_k^I) \quad (3.10)$$

where the jet similarity \mathcal{S} is either \mathcal{S}_{abs} or \mathcal{S}_{pha} (see equations 3.7 and 3.8).

The process of graph matching is divided into two stages. In the first stage the graph is shifted across the image while keeping its form rigid. Steps of one pixel in either direction are used for this rigid shift. For each position of the graph jets are extracted from the image at the vertex positions and the total similarity of the newly positioned graph \mathcal{G}^I to the original graph \mathcal{G} is calculated. This *global move* is able to position the graph on the object. The position with the highest similarity is the starting position for the second stage, which permits small graph distortions, i.e., the vertices are shifted locally and independently a small distance from their starting position. A region of 5×5 pixels is chosen for each vertex, which is scanned in steps of one pixel in either direction. Again, jets are extracted and the total similarity is calculated for each step. After this *local move* the optimal position of the graph is found at the position which provides the highest total similarity (see figure 3.8 for an example).

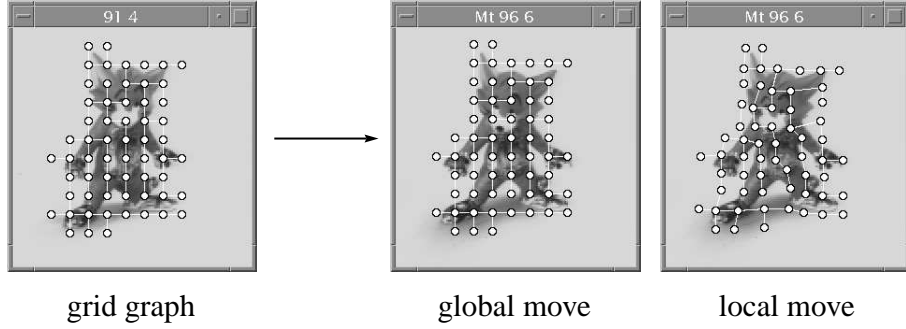


Figure 3.8: Matching Local Object Features. The grid graph which represents view (91, 4) is matched on view (96, 6). After the global move the rigid graph has found its optimal position on the object. After the local move the graph has been deformed to optimize the positions of the single vertices.

3.7 Tracking Local Object Features

In this section another technique is described which is used to measure the similarity between neighboring object views. In addition, it provides pixel positions of corresponding object points for different views. The *tracking* of local object features along a sequence of a moving or rotating object is described by Maurer and von der Malsburg [36] and based on an idea of Fleet and Jepson [18] and Theimer and Mallot [71]. Given a sequence of a rotating object and the pixel position of an object point for view r of the sequence, the aim is to find the corresponding position of the object point in view $r + 1$. As a visual feature Gabor wavelet responses are used again. For tracking a local feature from one view to another, a similarity function between two jets \mathcal{J} and \mathcal{J}' is defined, which differs slightly from the similarity function 3.8:

$$S_{tra}(\mathcal{J}, \mathcal{J}', \vec{d}) = \frac{\sum_i a_i a'_i \cos(\phi_i - \phi'_i - \vec{d} \vec{k}_i)}{\sqrt{\sum_i a_i^2 \sum_i a_i'^2}} \quad (3.11)$$

with \vec{d} being the displacement vector of the two jets and \vec{k}_i being the wave vectors of the Gabor filters. If \mathcal{J} and \mathcal{J}' are extracted at same pixel positions in the views r and $r + 1$, \vec{d} (and thus the new position of the object point) can be found by maximizing \mathcal{S} in its Taylor expansion with respect to \vec{d} . Because the estimation of \vec{d} is precise for a small displacement only, i.e., a large overlap of the Gabor jets, large displacement vectors are treated as a first estimate only and the process is iterated. Four iterations are used. In this way displacements up to half the wavelength of the kernel with the lowest frequency can be computed (see [82] for details).

For each vertex of the graph of view r the displacements are calculated for view $r + 1$. Then a graph is created with its vertices at the new corresponding positions in frame $r + 1$, and the labels of the new vertices are extracted from the new positions. But although the displacement vectors have been determined as decimal numbers, the jets can be extracted at (natural number) pixel positions only. This would result in a systematic rounding error. To compensate for this subpixel error $\Delta \vec{d}$ the phases of the Gabor filter responses



Figure 3.9: Tracking Local Object Features. These results taken from Maurer and von der Malsburg [36] show the views 1, 10, 20, 30, 40, and 50 of a tracked sequence.

are shifted according to $\Delta\phi_i = \Delta\vec{d} \cdot \vec{k}_i$. Then they take on values very similar to those that would be extracted at the correct subpixel positions (see figure 3.9).

Chapter 4

Robustness of Views Against Pose Variation

In this chapter the question will be treated in which direction and to which extent a three-dimensional object can be rotated so that no big differences occur between the starting view and the new views arising by rotation. I begin in section 4.1 by defining for a given view a surrounding area of viewpoints called *view bubble*, which is an essential part of this thesis. Inside its view bubble a given view is robust against pose variation. The remaining sections of this chapter deal with the question of choosing a suitable technique to determine view bubbles for a concrete object. For that purpose the matching and the tracking techniques introduced in sections 3.6 and 3.7 are compared with respect to essential requirements on view bubbles.

4.1 View Bubbles - A Measure of Pose Robustness

If a three-dimensional object should be represented sparsely by only some two-dimensional views of it - which is the main goal of this thesis - it is reasonable to choose such views which are representative for an area of viewpoints as large as possible. On the other hand, as object views which will not be chosen for the representation should be recoverable by interpolation it is necessary that the positions of corresponding object points in the chosen views are known. To facilitate an advantageous selection of views for the object representation first a surrounding area of robustness against pose variation is determined for each view. This area is called *view bubble*. The view bubble of view (p, q) is defined as the largest possible surrounding area of views on the viewing hemisphere for which the following two conditions hold:

- c1** The views constituting the view bubble are similar to view (p, q) .
- c2** Corresponding object points are known or can be inferred for each view of the view bubble.

These conditions guarantee the robustness of view (p, q) against pose variation within its bubble.

An ideal view bubble, which meets both conditions, may have an irregular shape. To simplify the determination of a view bubble I approximate it by a rectangle with view

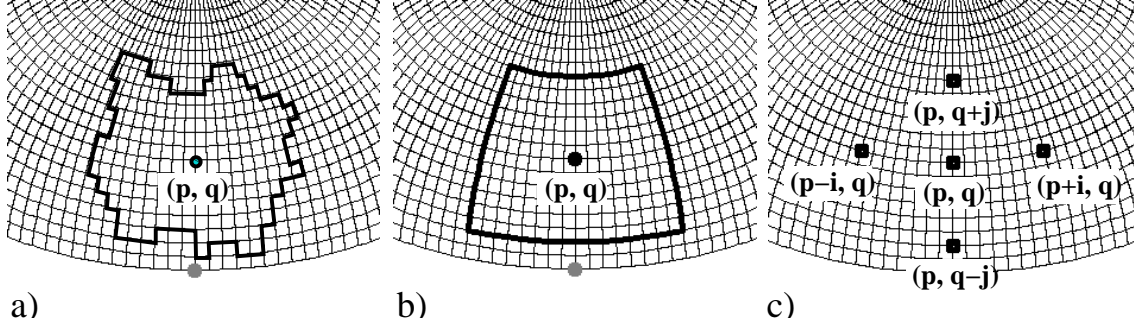


Figure 4.1: View Bubble and its Approximation. The ideal view bubble for a view (p, q) might have an irregular shape as depicted in part a). It can be approximated by a rectangle as shown in part b). Four border views define the rectangle (part c)).

(p, q) in its center (see figure 4.1 a) and b)). In the following the term *view bubble* is also used for this rectangular area. The set of views enclosed by the rectangle is denoted by $B_{(p,q)}$.

To determine the view bubble $B_{(p,q)}$ for view (p, q) , first views on the line of latitude of view (p, q) are compared. I begin with the views $(p-1, q)$ and $(p+1, q)$, taking the wrap-around topology of the viewing hemisphere into account. If both views provide a sufficiently high similarity to view (p, q) the algorithm goes on with the views $(p-2, q)$ and $(p+2, q)$. This procedure is stopped if one of both tested views becomes too dissimilar to view (p, q) . For the line of longitude this procedure stops for the same condition and in addition it stops if the top view for $q = 24$ or the bottom view for $q = 0$ is reached. Thus, for each view (p, q) on the hemisphere four *border views* $(p-i, q)$, $(p+i, q)$, $(p, q-j)$, and $(p, q+j)$ are obtained which provide a sufficiently high similarity to the center view (p, q) and which define the view bubble $B_{(p,q)}$ for it (see figure 4.1 c)). The similarity between views is measured by using similarity function 3.10 with either the jet similarity \mathcal{S}_{abs} or \mathcal{S}_{pha} . In this chapter function \mathcal{S}_{pha} is used, whereas in the following chapters the generation of view bubbles is performed using the coarser function \mathcal{S}_{abs} , because it showed itself to provide sufficient results.

Because each view (p, q) is represented by a labeled graph $\mathcal{G}_{(p,q)}$, I define the *representation of the view bubble* $B_{(p,q)}$ of view (p, q) by the graphs of the center view and the four border views and denote it by $\mathcal{B}_{(p,q)}$:

$$\mathcal{B}_{(p,q)} := \langle \mathcal{G}_{(p,q)}, \mathcal{G}_{(p,q)}^w, \mathcal{G}_{(p,q)}^e, \mathcal{G}_{(p,q)}^s, \mathcal{G}_{(p,q)}^n \rangle \quad (4.1)$$

The letters w , e , s , and n stand for *west*, *east*, *south*, and *north*. Graph $\mathcal{G}_{(p,q)}$ of the center view is the graph which is generated from the original image $I_{(p,q)}$ as described in section 3.5, whereas the graphs $\mathcal{G}_{(p,q)}^w, \mathcal{G}_{(p,q)}^e, \mathcal{G}_{(p,q)}^s$, and $\mathcal{G}_{(p,q)}^n$ are derived from graph $\mathcal{G}_{(p,q)}$ by either matching or tracking its jets to the border views. $\mathcal{S}_{tot}(\mathcal{G}_{(p,q)}, \mathcal{G}_{(p,q)}^w) \geq \tau$ holds for a preset similarity threshold τ (see equation 3.10). Accordingly, the same inequality with the same threshold τ holds for the east, south, and north graphs as well.

To determine $\mathcal{B}_{(p,q)}$ for a given view (p, q) by *matching*, graph $\mathcal{G}_{(p,q)}$ is matched successively in the neighboring views on the lines of latitude and longitude as described in section 3.6. The total similarity $\mathcal{S}_{tot}(\mathcal{G}_{(p,q)}, \mathcal{G}_{(p',q')})$ of the center graph to the graph ex-

tracted from the new view (p', q') is calculated for each new tested view using similarity function \mathcal{S}_{pha} (equation 3.8).

In this chapter smaller grid graphs are used as those described in section 3.5. They are generated by deleting all vertices which lie on the background or which lie on the object but are too close to the background. The minimal allowed distance of an object vertex to the background segment is a fraction of 10% of the width σ/k_ν of the Gaussian of the largest Gabor kernel. The reason for this is to prevent vertices from incorporating too much information on the background. Another drawback of larger graphs is the possibility of tracked vertices clinging to the outline of the object while it rotates. This could be disadvantageous especially if the quality of the tracking procedure is to be analyzed. All further studies outside this chapter are carried out utilizing the larger grid graphs introduced in 3.5 because they incorporate more information about the object.

If the view bubble $\mathcal{B}_{(p,q)}$ for view (p, q) should be determined by *tracking* I start with graph $\mathcal{G}_{(p,q)}$, which is tracked to all directions (west, east, south, and north) as described in section 3.7. For this chapter, after each tracking step, i.e., after each new tested view, the same total graph similarity is computed as for the matching procedure. The difference between tracking and matching lies in the fact that during matching each view is treated independently, whereas the tracking procedure utilizes the continuity of neighboring views. This leads to the question which method is more appropriate to determine areas of pose robustness for views of real objects. The remaining sections of this chapter deal with this question and have already been reported by Peters et al. [55].

4.2 Methods of Comparing Matching With Tracking of Local Object Features

The defining conditions **c1** and **c2** of a view bubble provide two criteria to judge the appropriateness of the matching and tracking procedure for the generation of view bubbles. To compare the similarity measure of both procedures one has to look at the sizes of the view bubbles. For a preset similarity threshold the sizes of the view bubbles resulting from matching and tracking are compared. This quantitative criterion is described in subsection 4.2.1. The second criterion consists in a qualitative assessment of the correspondences provided by both methods. This is described in subsection 4.2.2. For both procedures the same similarity threshold $\tau = 0.77$ has been used. (The average similarity between two randomly chosen views of the “dwarf” object was 0.68, determined for 15 matched pairs of views.)

4.2.1 Quantitative Comparison

For the quantitative comparison I determined the *view bubble area* for each view on the hemisphere by calculating the area of the approximating rectangle which is defined by the four border views of the view bubble, i.e., the view bubble area is $4 \cdot i \cdot j$ (see figure 4.1). Simulations have been made for both objects and both procedures, matching and tracking. For each object a *t*-test has been carried out to prove the hypothesis of different means

of the areas of view bubbles for the samples “view bubbles generated by matching” and “view bubbles generated by tracking”.¹

4.2.2 Qualitative Comparison

For the qualitative comparison four sequences of successive views on the hemisphere (from a starting view to a destination view) were chosen for both objects. For each of these sequences the matching and the tracking procedures have been performed, respectively. The sequences were selected arbitrarily and their length was determined by the last tracking step which provided good correspondences. They have an average length of about 8 views, which means they cover a rotation angle of 25.2 degrees. The longest sequence covers 43.2 degrees and consists of 13 views. To assess the correspondences by visual inspection, the resulting matched and tracked graphs are displayed for each view and the calculated similarities are plotted in diagrams. The results are described in subsection 4.3.2.

4.3 Results

4.3.1 Quantitative Comparison

Diagrams which show the distributions of view bubble areas are depicted in figure 4.2 for the “Tom” object and in figure 4.3 for the “dwarf” object. In both figures the first diagram shows the results from the tracking procedure and the second diagram the ones from the matching procedure. Light colors encode large areas, dark colors encode small areas of view bubbles. To compare the results for tracking and matching the third diagram shows the difference between the first and second diagram. Dark areas in the third diagram are areas on the hemisphere where tracking provides larger view bubbles than matching.

From the diagrams the following results can be derived. The tracking procedure provides larger view bubbles than the matching procedure for the majority of views for the more complex “Tom” object, whereas for the simpler “dwarf” object it is the other way around. Here the matching procedure provides larger view bubbles than the tracking procedure for the majority of views. The one-tailed t -test, with which the mean values have been compared, was significant with $\alpha = 1\%$ for each case. Some statistical values are summarized in the tables 4.1 and 4.2.

¹In a report by Peters et al. [54] another criterion for a quantitative comparison is described. For each view it is counted in how many other view bubbles it is contained. This criterion provides results similar to those reported in this chapter.

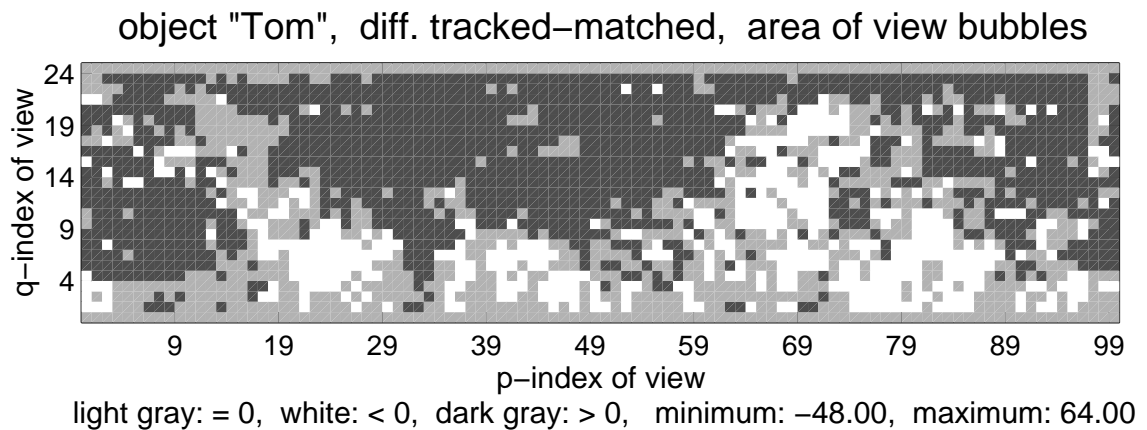
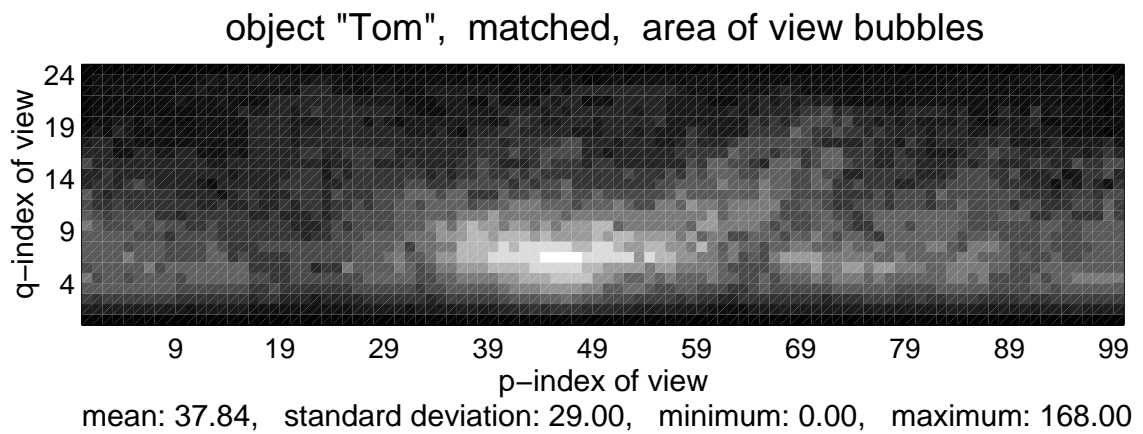
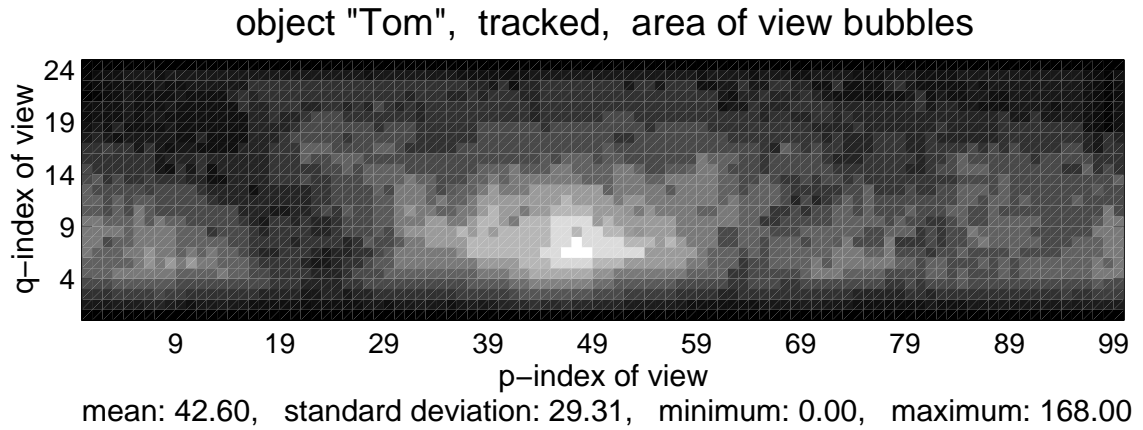


Figure 4.2: Object "Tom", Area of View Bubbles.

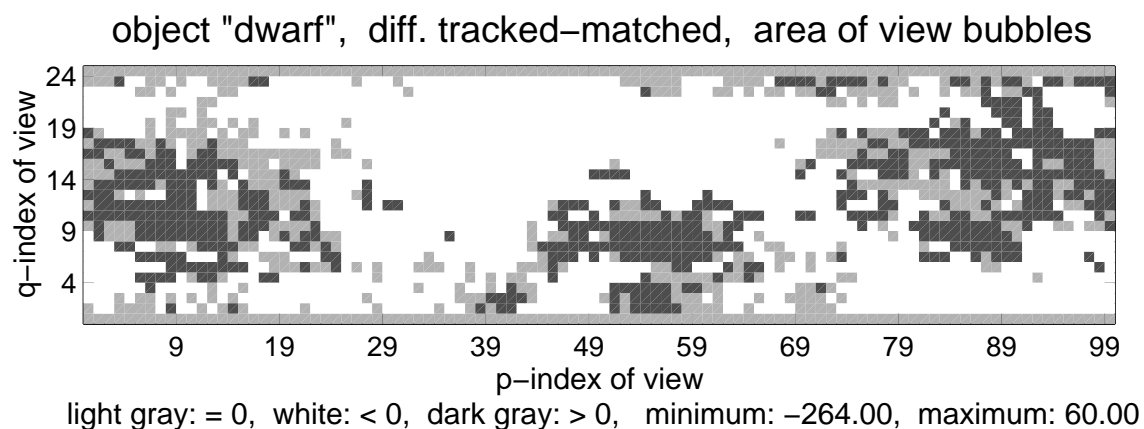
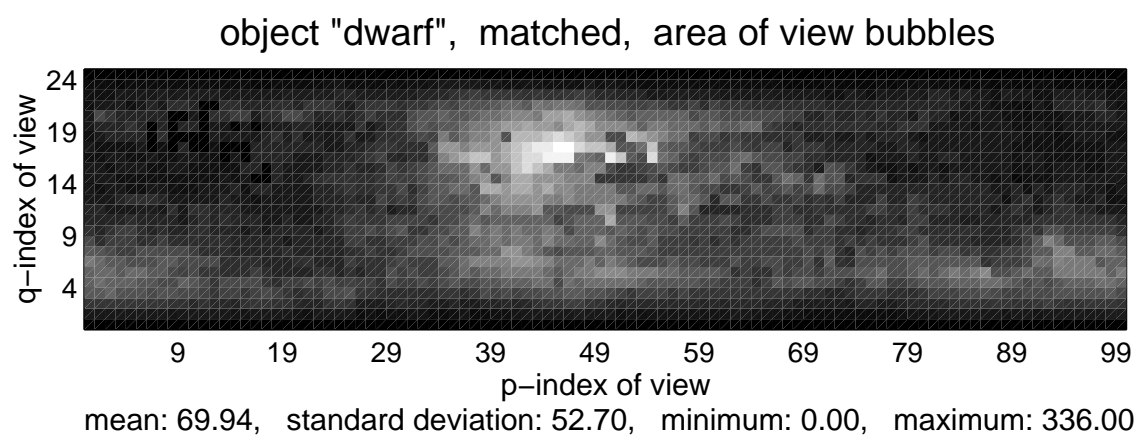
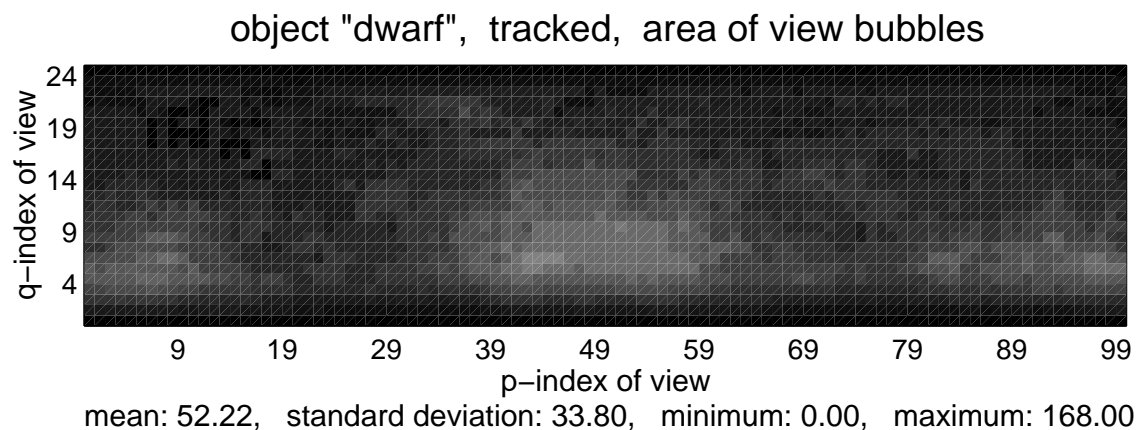


Figure 4.3: Object "Dwarf", Area of View Bubbles.

object “Tom”, area of view bubbles			
tracking		matching	
mean	= 42.60	mean	= 37.84
standard deviation	= 29.31	standard deviation	= 29.00
minimum value	= 0.00	minimum value	= 0.00
maximum value	= 168.00	maximum value	= 168.00
t -test with $\alpha = 1.0\%$: $T = 5.77 \implies \text{mean}_{\text{track}} > \text{mean}_{\text{match}}$			

Table 4.1: Statistics for Object “Tom”

object “dwarf”, area of view bubbles			
tracking		matching	
mean	= 52.22	mean	= 69.94
standard deviation	= 33.80	standard deviation	= 52.70
minimum value	= 0.00	minimum value	= 0.00
maximum value	= 168.00	maximum value	= 336.00
t -test with $\alpha = 1.0\%$: $T = 14.16 \implies \text{mean}_{\text{match}} > \text{mean}_{\text{track}}$			

Table 4.2: Statistics for Object “Dwarf”

4.3.2 Qualitative Comparison

In figure 4.5 the results for one sequence of the “Tom” object are shown, in figure 4.6 for one sequence of the “dwarf” object. The results for the other sequences can be found in appendix A.

In the first part of each figure the views of the object with the graphs resulting from the tracking procedure (first row of images) and the views of the object with the graphs resulting from the matching procedure (second row of images) are displayed. Both rows start with the starting view of the sequence with its original grid graph on the object. The next two images are chosen according to the quality of the matching, which is measured by visual inspection. The view with its graph where the matching provided the last successfully matched graph of the sequence is shown. The subsequent image depicts the view with the first mismatched graph of the sequence. Arrows point to the mismatched vertices. The last images of the rows show the last views of the sequence where the tracked graph still keeps the corresponding points, whereas the matched graph does not. In the headers of the images the indices of the views can be read. “Tr” means “tracked”, “Mt” means “matched”. These images only show a part of the complete sequences, the complete sequences with their tracked and matched graphs are depicted in appendix A as well.

The second part of each figure shows a diagram where the similarity of each view of

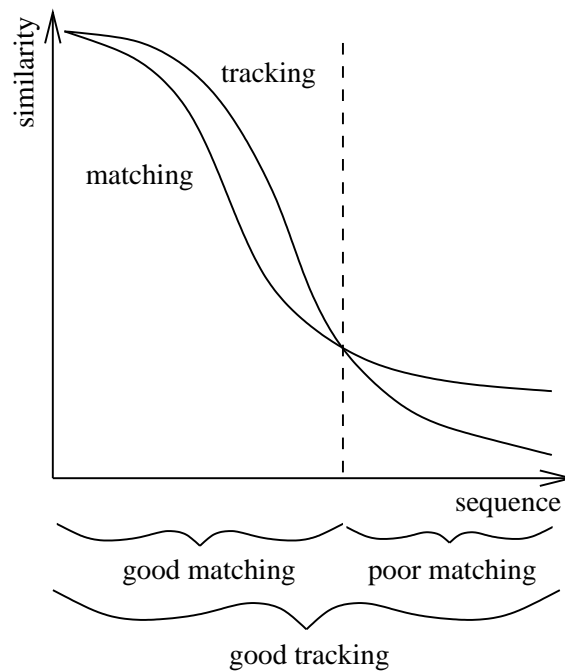


Figure 4.4: Qualitative Similarity Diagram. “Good” and “poor” is meant in the sense of correct, respectively incorrect, correspondences. See description in the text for details.

the sequence to the starting view is plotted for the tracked as well as the matched graphs. The similarities decrease monotonously while the object rotates away from the starting view, for the tracking as well as for the matching procedure.

From the assessment of the positions of the vertices of the tracked and matched graphs I can make the statement that for each view of each sequence the tracking procedure provides the same or better correspondences than the matching procedure. For the last view of each sequence the tracking procedure provides considerably better correspondences than the matching procedure.

From the similarity diagrams I get the following result. At the beginning of a sequence the tracking procedure always provides slightly higher similarities than the matching procedure. This relationship is reversed at that point of the sequence where the matching starts to provide poor correspondences, whereas tracking provides good correspondences until the end of the sequence (see figure 4.4).

TOM, sequence (48,5) -> (36,5)

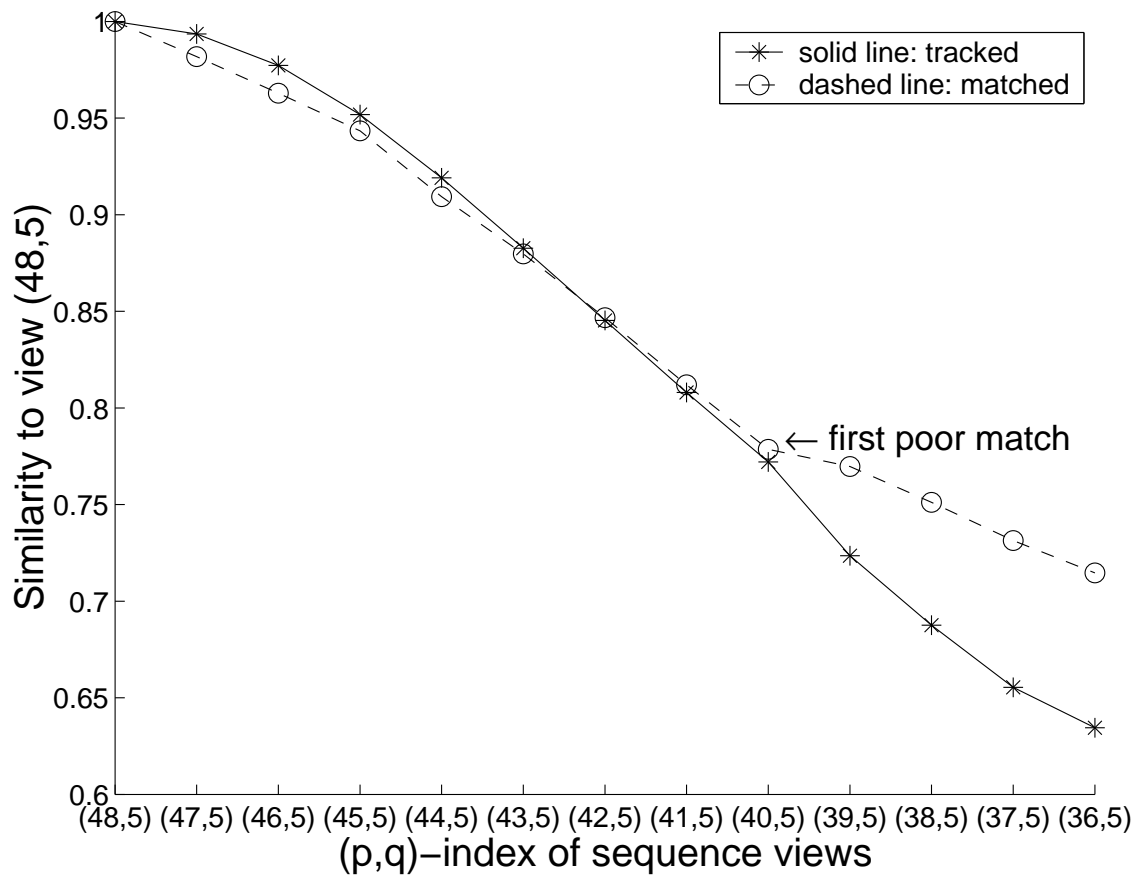
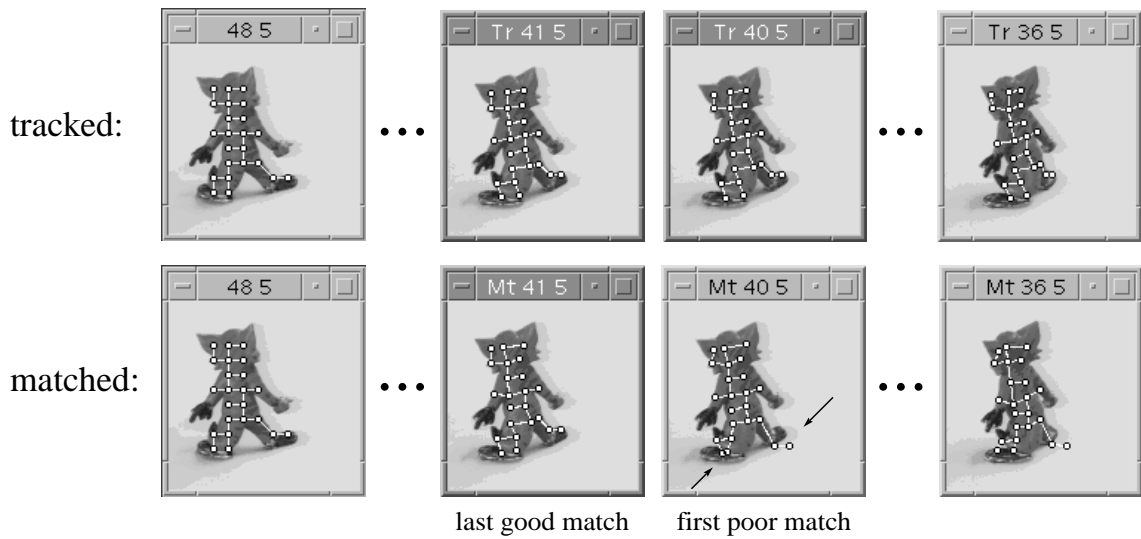


Figure 4.5: Object “Tom”, First Sequence With Similarity Diagram.

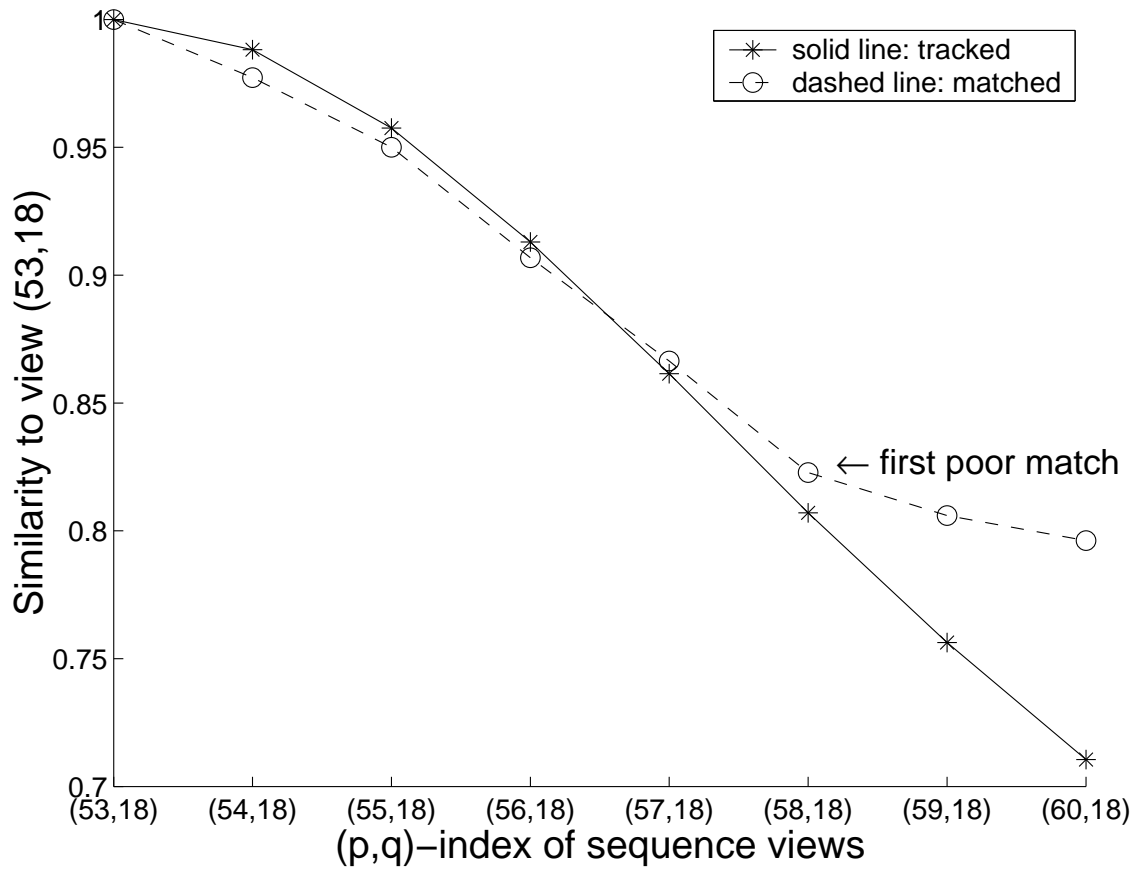
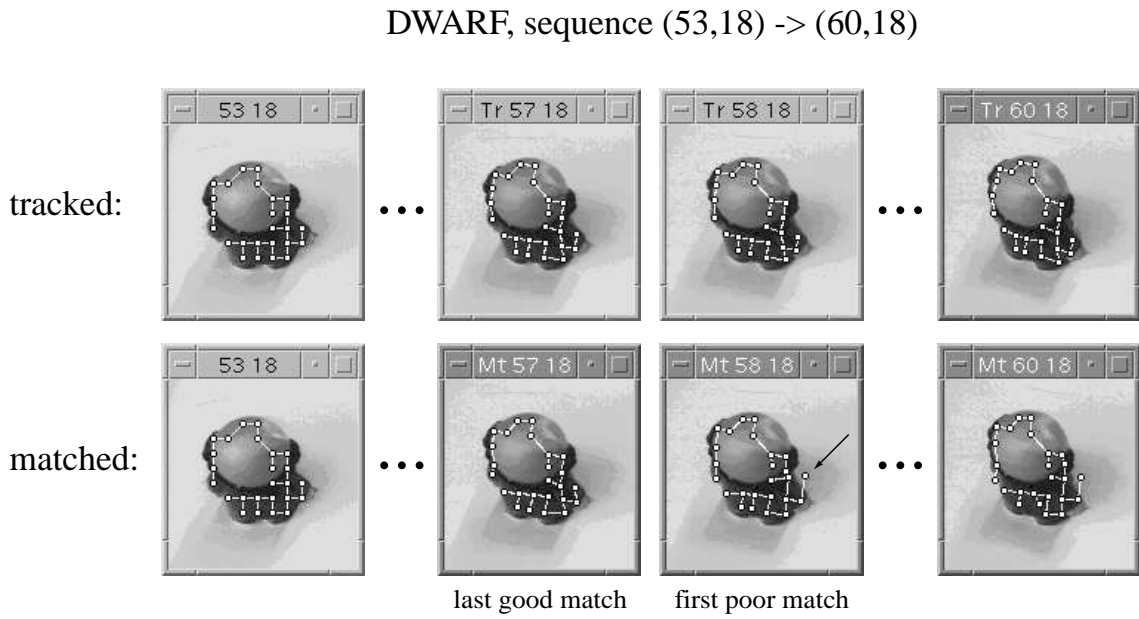


Figure 4.6: Object “Dwarf”, First Sequence With Similarity Diagram.

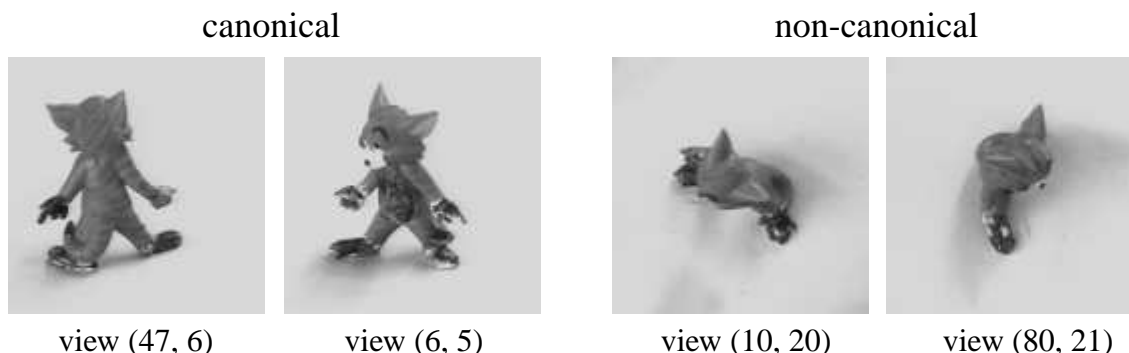


Figure 4.7: Canonical and Non-Canonical Views for Object “Tom”. View (47, 6) is the view which provides the global maximum of its view bubble area (generated by the tracking procedure). Its view bubble covers an angle of 50.4 degrees on the line of latitude and 43.2 degrees on the line of longitude. Thus, view (47, 6) can be regarded as canonical view. The same is true for view (6, 5). By contrast, the views (10, 20) and (80, 21), for example, provide very small view bubbles only and they can be declared as non-canonical views. Please, compare these views with the first diagram of figure 4.2, where the canonical views are located in light areas whereas the non-canonical views are located in dark areas.

4.3.3 Canonical Views

Besides the differences between the distributions of view bubble areas for the “Tom” object and the “dwarf” object, respectively, which were described in subsection 4.3.1, they have something in common. For both objects the distribution of view bubble areas is qualitatively similar for the tracking procedure and for the matching procedure. The back view seen from slightly above and the front view provide the largest bubbles, and they can be regarded as canonical views as they have been introduced in subsection 2.1.3. In general, a view can be defined as canonical view, if its view bubble area is a local maximum and if it is larger as a preset threshold (see figure 4.7 for examples).

4.4 Discussion

In this chapter two procedures for finding areas of pose robustness on the viewing hemisphere of a three-dimensional object have been compared. For each view on the viewing hemisphere a view bubble is defined, which is a surrounding area of robustness against pose variation. During the investigations the question was pursued whether one of the procedures, tracking or matching of local object features, outperforms the other in the determination of view bubbles. At first glance both procedures are suitable for this task. Areas of large and small view bubbles arise on the viewing hemisphere. Views in the centers of areas of large view bubbles are more robust against pose variations than other views and can be regarded as *canonical views*. This result provides an answer to the questions Q2 formulated in the introduction (chapter 1).

As pose robustness is defined by a quantitative and a qualitative property (the *largest* area which preserves *correspondences*) I have two comparison criteria. With regard to

the quantitative criterion I have detected no difference between matching and tracking concerning the size of the view bubbles given a fixed similarity threshold. But with regard to the qualitative criterion I found that much more precise correspondences were provided by tracking than by matching.

In detail, from both test objects no statement was possible about the superiority of one procedure in terms of the size of the view bubbles, because for the more complex “Tom” object tracking provided larger view bubbles, whereas matching outperformed tracking for the simpler “dwarf” object. (For the difference between both objects refer to section 3.1.) A possible explanation for this result could be that the rapidly changing views of the “Tom” object cannot be matched over larger distances, because the matching procedure is looking for the *same* appearance of the object features, whereas the tracking procedure reacts more sensitively when views are changing during the rotation of the object. A hypothesis derived from the statistics is that the tracking procedure leads to larger view bubbles than the matching procedure for “complex” objects, whereas matching is superior to tracking for “simple” objects. This hypothesis should be verified by analyzing more examples. The minimum size of view bubble areas is zero for both objects and both procedures. It occurs for views in the north pole and on the equator, because for those views $j = 0$ is fulfilled. Although this is an artefact of the proposed method I suppose that it has no effect on the comparisons, because it occurs for both procedures and both objects alike.

The results of the qualitative comparison have been obtained from the analysis of the test sequences. These sequences have been chosen arbitrarily and are located in different positions on the viewing hemisphere. They display the objects from front, back, and side views and the object points are tracked in different directions. Thus, I assume that the results obtained from these test sequences are representative and can be generalized to any sequence on the viewing hemisphere. A reason for the more precise correspondences found by tracking could be the fact that an object feature changes its appearance while the object rotates. The feature in the tracking procedure adapts to this change, whereas the matching procedure always searches for the same starting feature. The more the rotation proceeds the more difficult it is for the matching procedure to find the correct point. The advantage of tracking of object features is that it “joins in” the rotation. *Continuous information* is used by the tracking procedure in contrast to the matching procedure. Matching is the more appropriate method if the task is to find features with the *same* appearance, tracking is the more appropriate method if *changes* of the features should be followed.

A disadvantage of the automatical processing of the images is that in some cases a relatively poor result of the segmentation of the images leads to a representing grid graph which does not cover the whole object (as, for example, in figure 4.6) or which covers larger parts of the background. But as these problems occur for both compared procedures alike, I claim that they did not influence the results.

Even if it turns out that matching is superior to tracking for simple objects in terms of the size of the view bubbles I consider the qualitative requirement more important. Precise correspondences should take priority over larger view bubbles, particularly for further processing. For the generation of unfamiliar views, for example, as described in the following chapters 6 and 7, precise correspondences in the familiar views are necessary, and to establish these correspondences the continuity information of successive views has

to be utilized. Accordingly, my conclusion for this chapter is that the tracking of object features is superior to matching for the determination of areas of pose robustness for a three-dimensional object, especially for complex objects. Thus, for the remaining chapters of this thesis view bubbles are determined by tracking. Coming back to formula 4.1 this means that the graphs $\mathcal{G}_{(p,q)}^w$, $\mathcal{G}_{(p,q)}^e$, $\mathcal{G}_{(p,q)}^s$, and $\mathcal{G}_{(p,q)}^n$ are derived from graph $\mathcal{G}_{(p,q)}$ by *tracking* its jets to the border views.

4.5 Parallels to Primate Object Perception

My data suggest that the tracking procedure provides more precise correspondences in neighbouring views of a three-dimensional object than the matching procedure. In other words, good correspondences are derived from the *continuity* of successive views and not from *disconnected* static views. This result is supported by the research of P. J. Kellman [25]. His experiments with infants suggest that they have the ability to perceive the three-dimensional form of an object only if information about continuous optical transformations given by motion is available. They are not able to apprehend the overall form of an object from static views, even if they are multiple or sequential. This holds for even eight months old infants. Adults, however, are able to perceive a three-dimensional form from static views, as already mentioned in subsection 2.2.2.

That context in general (not necessarily temporal) can improve the recognition of novel views was shown by Christou et. al. [12]. However, temporal context seems to be of special importance for object perception. Numerous psychophysical experiments support this. For example, T. Niemann et al. [44] report on experiments with parts of statues of human figures on a turntable. They recorded the eye movements of subjects watching the rotating objects and found the eye movements often directed to the same details seen from different vantage points. This also supports the relevance of tracking of local features.

Another argument is furnished by K. L. Harman and G. K. Humphrey [21]. They claim that different object representations are generated, depending on the presentation of either regular or random sequences of views of the object. When a sequence of rotation is encoded, the associated temporal context may lead to the construction of a linked, higher-order system of representations for a given object, whereas, without temporal context, a single representation of each object view may be constructed.

Also some physiological reasons emphasize the importance of the successive appearance of views for the learning of an object representation. Miyashita [39] trained monkeys to match complex fractal patterns, which were presented successively in a fixed series of 100 items. After training some cells in the anterior temporal cortex were found to show selectivity for a small number of patterns which had been presented successively. This gives evidence for learning based on temporal associations rather than on pattern overlap.

Perrett et. al. [47] already observed that an object representation in the form of a collection of stored views is structured in the sense, that views belonging together because of their successive appearance are more closely associated with each other in the representation. This was confirmed in a later study by Edelman and Weinshall [17].

Chapter 5

Sparse Object Representation

“The knowledge of only some views of a three-dimensional object is sufficient to solve perception problems such as estimating the pose of the object.” As this is the main thesis of this work, it is time now to think about the composition of an object representation which meets these requirements. As already mentioned in the introduction (chapter 1) the following conditions should hold for such a representation of a three-dimensional object:

1. It should be constituted from *two-dimensional views* of the object.
2. It should be *sparse*, i.e., it should consist of representations of *as few views as possible*.
3. It should be capable of *performing perception tasks*.

In section 5.1 I propose an object representation which meets these conditions. Section 5.2 gives some examples of sparse representations for concrete objects. The sparseness of the proposed representation is checked in the chapters 6 and 7, whereas the third condition is examined in chapter 8.

5.1 Generation of a Sparse Object Representation

In the last chapters the representation of a single object view by a labeled graph (section 3.5) and the assignment of an area of pose robustness (view bubble) to each view on the viewing hemisphere (section 4.1) have been described.

To meet the first two conditions of a representation the aim is now to choose single object views (in the form of labeled graphs) to constitute the representation. The distribution of the view bubbles can function as a decision criterion. Starting from this distribution it is obvious that having one view bubble for each view on the viewing hemisphere they overlap each other to a large extent (see figure 5.1). The idea is to reduce this large number of view bubbles and to choose as few of them as possible which nevertheless cover the whole viewing hemisphere.

For the selection of the view bubbles I use a *greedy set cover algorithm* proposed by Chvatal [13], which is described in subsection 5.1.1. The set cover algorithm will provide a set of view bubbles which covers the whole viewing hemisphere. Its representation \mathcal{R} will then constitute the sparse object representation.

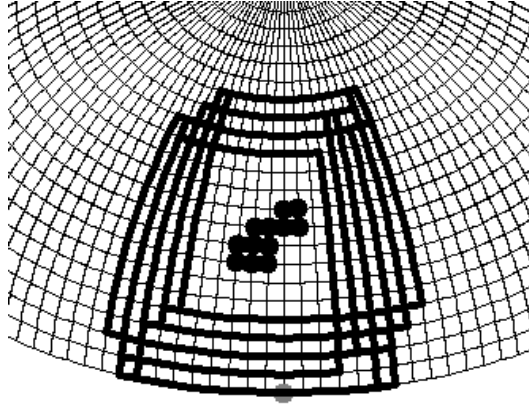


Figure 5.1: Overlap of View Bubbles. As a view bubble is determined for each view on the viewing hemisphere, they are overlapping on a large scale. This figure shows a part of the hemisphere with some view bubbles.

5.1.1 Set Cover Algorithm

In the set cover problem as it applies to this thesis the data consists of finite sets B_1, B_2, \dots, B_n . $\bigcup_{j \in J} B_j$ is denoted by H with $H = \{1, 2, \dots, m\}$, $J = \{1, 2, \dots, n\}$. A subset R of J is called a *cover* if $\bigcup_{j \in R} B_j = H$. The problem is to find a cover with a minimal number of elements. This problem is known to be NP-complete.

A binary matrix A of size $m \times n$ can be laid down where the columns encode the sets B_1, B_2, \dots, B_n , i.e., $a_{ij} = 1 \Leftrightarrow i \in B_j$. Then the sought subset R of J is the set of indices of chosen columns. A selection of columns can be expressed by ones in a binary vector $\vec{v} = (v_1, v_2, \dots, v_n)$. The problem consists in minimizing $\sum_{j=1}^n v_j$ subject to $\sum_{j=1}^n a_{ij} v_j \geq 1 \forall i = 1, \dots, m$. The last inequality guarantees the covering of all rows of matrix A , i.e., of all elements of H . The greedy heuristic algorithm proposed in [13] for solving the set cover problem does not necessarily find the true optimum but usually a feasible solution:

```

R = ∅
while (number of ones in matrix A > 0)
  find a column k with the largest number of ones
  R := R ∪ {k}
  set each row of matrix A to zero where column k has a one
end_while
R is a cover

```

Applied to the problem of covering the viewing hemisphere by view bubbles the sets B_1, B_2, \dots, B_n are the rectangular bubbles as introduced in section 4.1 with $n = 2500$ ¹. Accordingly, H is the set of all views, thus $m = 2500$ as well. J is the set of indices of the view bubbles. The above algorithm can now be expressed in the following way:

¹The double subscript (p, q) used there is reindexed here by a single subscript.

```

 $R = \emptyset$ 
while (there are still uncovered views on the hemisphere)
    find the view bubble which covers the largest number of uncovered views
    add this view bubble to the cover  $R$ 
    mark the views covered by this view bubble
end_while
 $R$  is a cover

```

The set R contains the indices of the view bubbles which cover the viewing hemisphere. I define now the *sparse representation* \mathcal{R} of a *three-dimensional object* as the set which consists of the representations $\mathcal{B}_i, i \in R$, of the view bubbles of set R :

$$\mathcal{R} := \{\mathcal{B}_i\}_{i \in R}. \quad (5.1)$$

As the representation of one view bubble consists of the graphs of its center and border views (see equation 4.1), the sparse object representation consists of a collection of graphs which represent single views of the object: $\mathcal{R} = \{\langle \mathcal{G}_i, \mathcal{G}_i^w, \mathcal{G}_i^e, \mathcal{G}_i^s, \mathcal{G}_i^n \rangle\}_{i \in R}$. Neighboring views of the sparse representation are “connected” by known corresponding object points (the correspondences between center and border views). The letter ν is used for the denotation of the number of view bubbles which constitute the sparse representation: $\nu = |\mathcal{R}|$. In the next section some examples of sparse representations of concrete objects are given.

5.2 Results

For both objects introduced in figure 1.3 view bubbles are generated for each view on the viewing hemisphere as described in the previous chapters. For each view graphs are used as described in section 3.5 and the view bubbles are generated by tracking using the jet similarity function \mathcal{S}_{abs} as described in section 4.1. After that the set cover algorithm described in the previous section is applied, resulting in a sparse object representation for each object. As the size of the view bubbles and the resulting object representation strongly depend on the similarity threshold τ for tracking (see section 4.1), I varied this threshold using five different values: 0.75, 0.8, 0.85, 0.9, and 0.95.

Figure 5.2 shows five different partitionings of the hemisphere for the “Tom” object depending on the value of the tracking threshold. For example, for $\tau = 0.75$ the set cover algorithm provides a cover consisting of 6 view bubbles. The diagrams show the covered hemisphere seen from above, the dots mark the center views of the cover bubbles. A larger threshold τ leads to smaller view bubbles, resulting in a cover with a larger number of bubbles which overlap to a smaller extent than for a smaller tracking threshold. Figure 5.3 shows the same for the “dwarf” object.

Each of the angle values given in the figures 5.2 and 5.3 indicates half of the width of view bubbles averaged over all bubbles of a cover, i.e., the average distance between the center and the east (or the west) views for all view bubbles. These values are summarized for both objects in table 5.1.

Object "Tom"

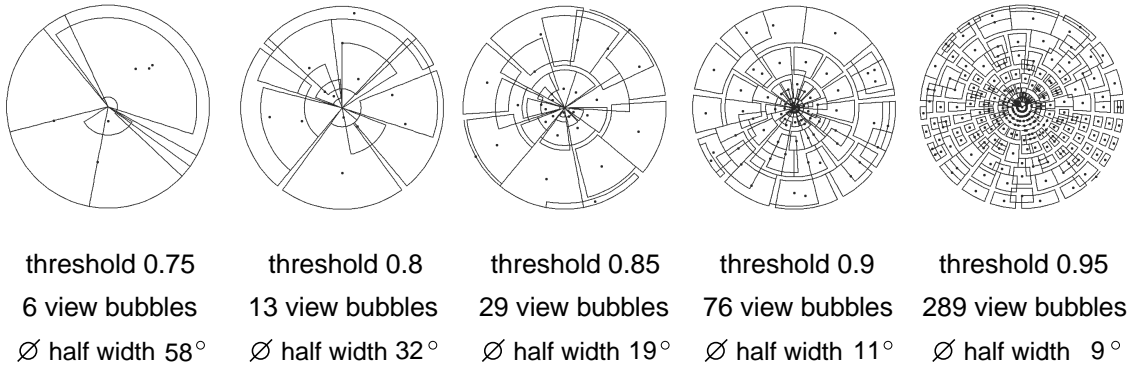


Figure 5.2: Five Different Covers for Object "Tom". See text for a description.

average angles between center and east views					
threshold	0.75	0.8	0.85	0.9	0.95
object "Tom"	58.29°	32.14°	19.32°	11.40°	8.84°
object "dwarf"	62.51°	38.64°	23.41°	13.32°	6.61°

Table 5.1: Average Angle Between Center and East View

Figure 5.4 shows that the algorithm provides similar but not the same results for both objects. Here the covered hemispheres are depicted from a different viewpoint.

Figure 5.5 shows for both objects the experimentally derived numbers ν of view bubbles plotted against the threshold τ used for generating the view bubbles. The same exponential fitting function

$$\nu = f(\tau) = \mathbf{r} \cdot e^{\mathbf{s} \cdot (\tau - \mathbf{t})}, \quad (5.2)$$

$\tau \in [0.75, 0.95]$, is drawn in the diagrams for both objects. It has been determined by using an iterative least squares algorithm for non-linear parameter fitting - the Levenberg-Marquardt method described by More [40] - for the "Tom" and "dwarf" data sets separately and then taking the average value of each fitted parameter, in this case of \mathbf{r} , \mathbf{s} , and \mathbf{t} . The data sets for both objects fit rather well to the fitting function, thus an exponential function seems to be an appropriate description of the correlation between the number of view bubbles and the similarity threshold. The slightly larger number of view bubbles for the "Tom" object for each similarity threshold can be explained by the fact that the "dwarf" object is a "simpler" object, whereas the views of the "Tom" object are changing

Object "dwarf"

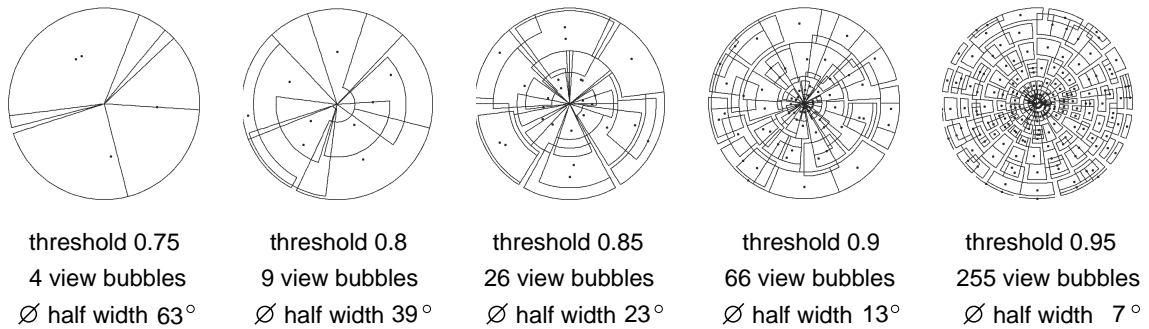


Figure 5.3: Five Different Covers for Object “Dwarf”. See text for a description.

more quickly. Thus more view bubbles are needed for the representation of “Tom”.

The graphs of the views which are depicted in figure 5.6 constitute the sparse representation for the “Tom” object for a tracking threshold of $\tau = 0.75$. The six view bubbles which constitute the representation are enclosed in boxes with their center and border views. Compare this figure with the first diagrams in the figures 5.2 and 5.4.

5.3 Discussion

The view bubbles which constitute the sparse representation for the medium partitionings of $\tau = 0.8, 0.85$, and 0.9 can be regarded as *aspects*, which have been introduced in subsection 2.1.2. Small changes of the viewpoint inside the range of a view bubble affect the appearance of the object only slightly, whereas the transitions between view bubbles are distinguished by qualitative changes in the object’s appearance, thus they can be regarded as *events*. This does not necessarily hold for the extreme covers of $\tau = 0.75$ and 0.95 . For $\tau = 0.75$ the constituting view bubbles are still overlapping to a large extent and cannot be regarded as distinct aspects separated by events. For the other extreme of $\tau = 0.95$ surely not each transition between neighboring view bubbles represents a significant change in the appearance of the object. Thus, I would only regard the medium partitionings as being convincing. Figure 5.6 demonstrates that the view bubbles constituting \mathcal{R} for $\tau = 0.75$ cannot be regarded as distinct aspects. Three of the six constituting view bubbles (b), c), and d)) are almost identical. This can be explained by the fact that none of them can be omitted to cover all views which are covered by the union of them. There is no smaller view bubble in the set of all view bubbles which would be able to cover the small number of views which would not be covered if one of the three bubbles in question was omitted.

The average distances between the center and the east (or the west) views listed in table 5.1 can be interpreted as average distances within which a generalization from the

hemisphere covering

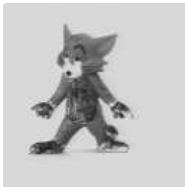
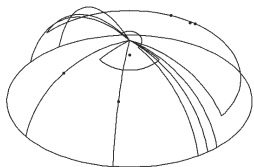
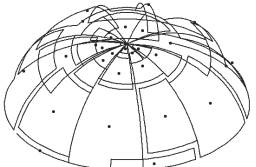
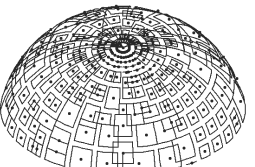

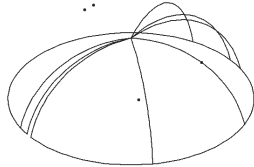
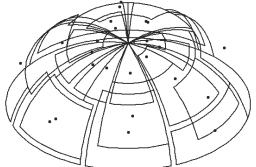
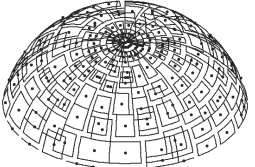
tracking threshold		0.75	0.85	0.95
object				
 Tom	 6	 29	 289	
 dwarf	 4	 26	 225	

Figure 5.4: Different Covers for Both Objects. The numbers next to the hemispheres are the numbers ν of view bubbles constituting the cover. See text for a description.

center view is possible. For $\tau = 0.8$ this distance lies *between 30° and 40°* , which is consistent with results from physiology and psychology as summarized in section 2.3 concerning questions **Q3**. Although similar for both test objects, these distances as well as the number and the distribution of the view bubbles seem to depend on the specific properties of the object.

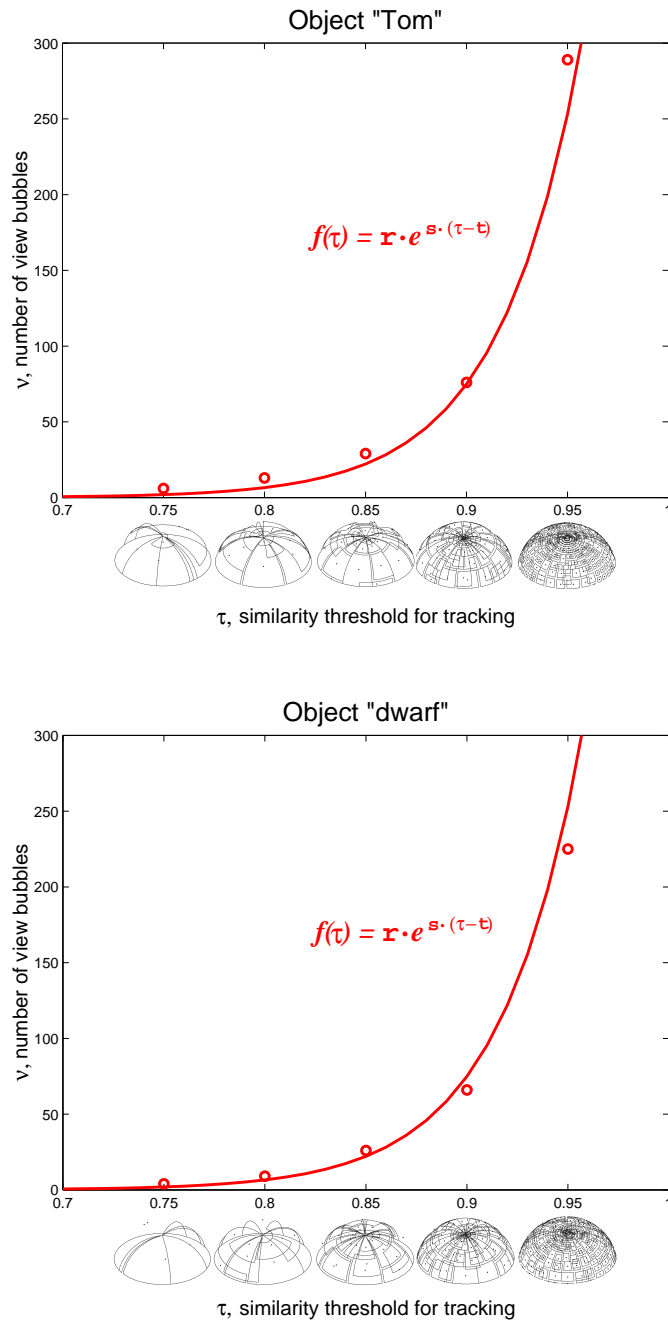


Figure 5.5: Correlation between Number of View Bubbles and Similarity Threshold. The small circles depict the measured values. The parameters of the fitting curve are calculated from the data sets for both objects and have the values $r = 926.297$, $s = 24.351$, and $t = 1.003$.

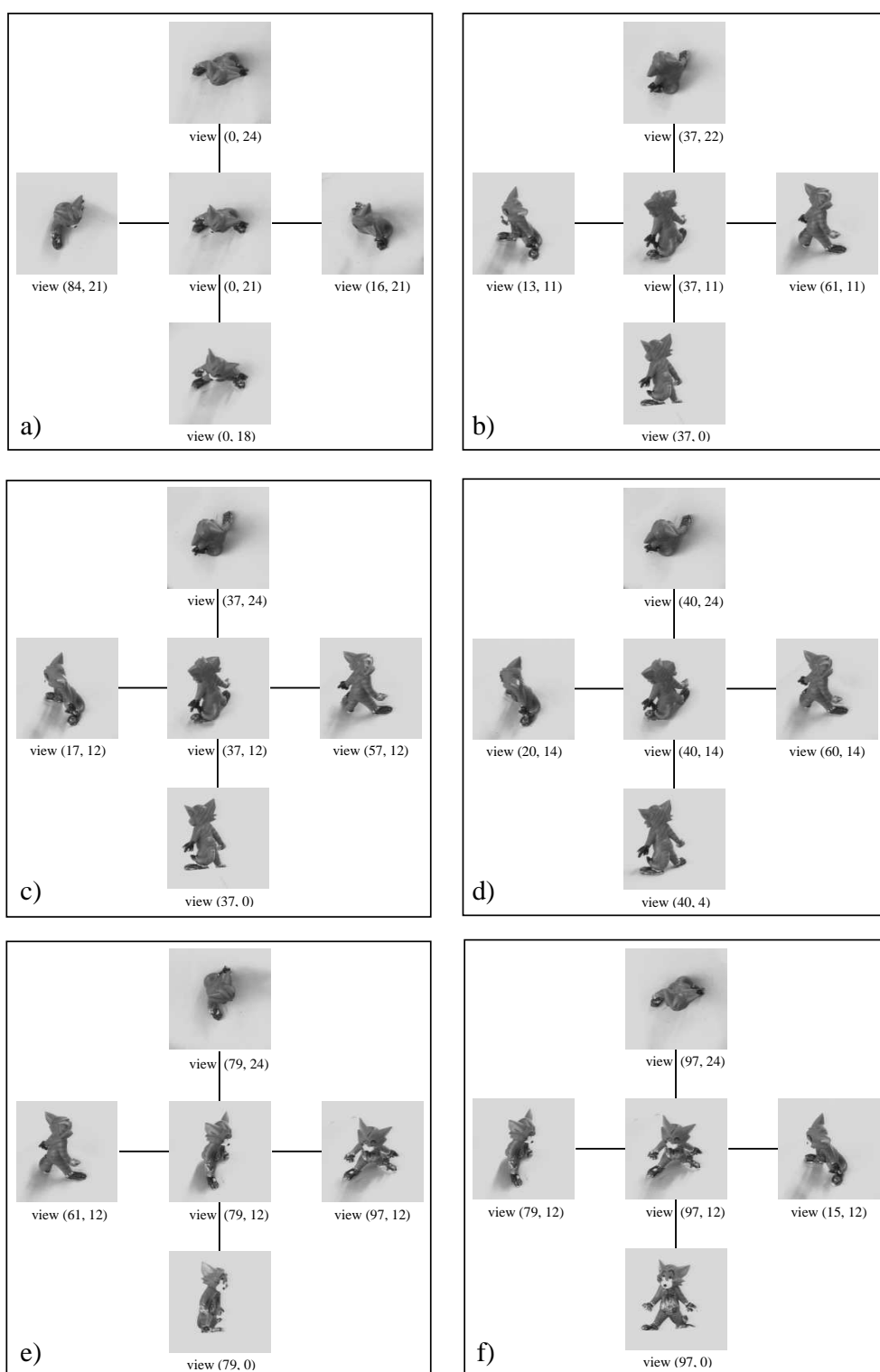


Figure 5.6: Sparse Representation of Object "Tom". The graphs which represent these views constitute the sparse object representation \mathcal{R} of the "Tom" object for a tracking threshold of $\tau = 0.75$. See text for more details.

Chapter 6

Morphed Views

The last chapter described how a view-based representation \mathcal{R} of a three-dimensional object can be obtained. One of the requirements claimed there was sparseness. In this chapter the sparseness of the representation is verified by checking the possibility to generate arbitrary, unfamiliar views of the represented object from the few views which constitute \mathcal{R} . Unfamiliar views are calculated from familiar views by a view morphing technique described in section 6.1. This requires the calculation of object point positions for unfamiliar views from available information about views stored in the representation. Object point positions for an unfamiliar view are determined by a linear combination of the corresponding point positions in views of the representation. As this linear combination of point positions is essential for techniques used in subsequent chapters as well, the morphing of object views also serves as a test for the quality of the linear combinations. View morphing as described in this chapter is not the main goal of this thesis, rather it serves as an auxiliary means and visualization tool. In section 6.2 the view morphing technique introduced in section 6.1 is assessed by an error analysis and a statistical description of a large set of morphed views.

6.1 Morphing of Unfamiliar Views

The term *morphing* originates from the word *metamorphosis* and describes smooth transitions between images as described by Beier and Neely [3, 65]. I will use this term for the transformation of sample views of the representation \mathcal{R} into unfamiliar views. As \mathcal{R} consists of graphs of center and border views of view bubbles which completely cover the viewing hemisphere it is possible to reconstruct any view from selected views of \mathcal{R} provided that the corresponding object points in the selected views are known. For that reason view morphing takes place exclusively *inside* a view bubble because only here the necessary correspondences between sample views are given.

Unfamiliar views of an object are computed from two as well as from three sample views. In the case of three sample views the unfamiliar view is calculated depending on its position in one of the four quadrants of the rectangle which defines the view bubble it belongs to. If the view to be generated lies in the first quadrant of its view bubble (including the axes) it is reconstructed from the center, the east, and the north view of the bubble, if it lies in the second quadrant it can be morphed from the center, the west,

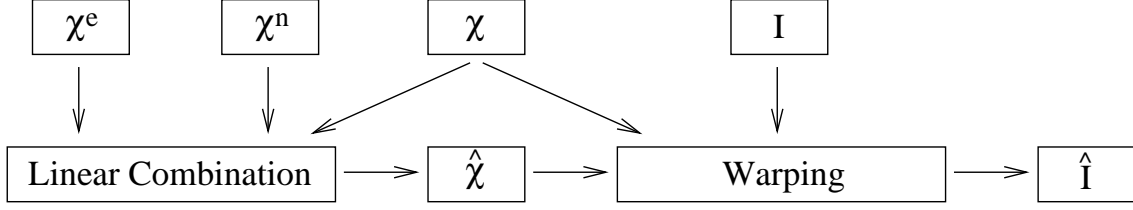


Figure 6.1: Flowchart of View Morphing From Three Sample Views. A novel view of an object is morphed in two steps. Firstly, the positions \mathcal{X} , \mathcal{X}^e and \mathcal{X}^n of object points in sample views are linearly combined (see equations 3.9 and 4.1). Secondly, the gray level image I of a sample view is morphed according to its own object point positions \mathcal{X} and the newly calculated positions $\hat{\mathcal{X}}$ resulting in the gray level image \hat{I} of the unfamiliar view. See also figure 6.2.

and the north view, and so on. If the unfamiliar view lies on the vertical axis of its view bubble the east view is always a source view, if it lies on the horizontal axis the north view is always a source view. In the case of two sample views the source views for morphing are either the center and the east view or the center and the west view of the bubble, depending on the position of the unfamiliar view to the right (including the axis) or the left of its view bubble's center view without regarding its vertical position in the view bubble.

View morphing proceeds in two steps (see figures 6.1 and 6.2):

1. In the first step *linear combinations* of the vertex positions of the graphs which represent the sample views are calculated. The resulting coordinates constitute the new vertex positions of the unfamiliar view to be reconstructed. For the example of three source views with the unfamiliar view lying in the first quadrant of its view bubble the problem can be formulated as follows:

given: the sets $\mathcal{X}_{(p,q)}$, $\mathcal{X}_{(p,q)}^e$, and $\mathcal{X}_{(p,q)}^n$ of object point positions in the sample views and the positions of the sample views and the unfamiliar view on the viewing hemisphere,

sought: the set $\hat{\mathcal{X}}_{(\hat{p},\hat{q})}$ of object point positions in an unfamiliar view (\hat{p},\hat{q}) which lies in the first quadrant of view bubble $B_{(p,q)}$.

For the case of other quadrants or only two sample views the given point sets vary accordingly. In subsection 6.1.1 a solution for this problem is proposed.

2. In the second step the given gray level image $I_{(p,q)}$ of the center view (p,q) is *warped* from its original object point positions $\mathcal{X}_{(p,q)}$ to the object point positions $\hat{\mathcal{X}}_{(\hat{p},\hat{q})}$ calculated in step one resulting in the final morphed image $\hat{I}_{(\hat{p},\hat{q})}$ of the unfamiliar view (\hat{p},\hat{q}) . This is described in more detail in subsection 6.1.2.

6.1.1 Linear Combination of Object Point Positions

In this subsection formulas are given for step one of the morphing algorithm. They have already been reported by Peters and von der Malsburg [53]. The development of the formulas was strongly motivated by Ullman and Basri's linear combination approach for

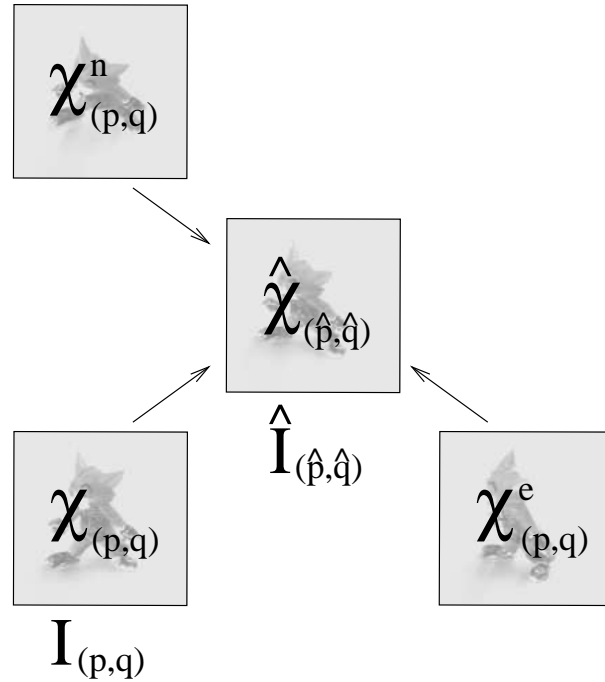


Figure 6.2: Morphing from Three Sample Views. See text and figure 6.1 for a description.

combining object views [74] already mentioned in subsection 2.1.4. Ullman and Basri proved theoretically the possibility to represent any unfamiliar, two-dimensional view of a three-dimensional object by a linear combination of a set of familiar, two-dimensional, sample views of the object under orthographic projection and provided that the same set of object points is visible in the sample views (consequently, no self-occlusions are allowed). As the application of their calculations to images of real objects bears some difficulties, they applied their algorithm to artificially created images, like line drawings of cars which consist of sets of two-dimensional contours. Difficulties of the application to real objects derive from the facts that the nature of the used object features has to be considered, that the correspondences in sample views have to be known, and that singularities due to self-occlusions can occur. Especially these visibility issues are crucial to understand *which* views can be synthesized. In my approach the correspondences are provided by the tracking algorithm and singularities are avoided by an appropriate choice of sample views. A detailed derivation of my formulas is given in appendix B. Under perfect conditions these formulas predict the same results for the view reconstruction from two as well as three sample views.

A view (p, q) can be identified with the pair (φ, λ) of its longitude and latitude angles on the viewing hemisphere with (i) $0 \leq \varphi < 2\pi$ and (ii) $0 \leq \lambda \leq \frac{\pi}{2}$ (see figure 6.3). Let $\varphi_1, \varphi_2, \varphi_3$, and $\hat{\varphi}$ denote the longitude angles and $\lambda_1, \lambda_2, \lambda_3$, and $\hat{\lambda}$ denote the latitude angles of the sample views and the unfamiliar view, respectively. The order of the numeration is chosen such that (iii) $\varphi_1 = \varphi_3$ and (iv) $\lambda_1 = \lambda_2$ hold. This is achieved by assigning index one to the center view of the view bubble and numbering the other views according to (iii) and (iv), e.g., view (φ_1, λ_1) is the center view, view (φ_2, λ_2) the east view and view

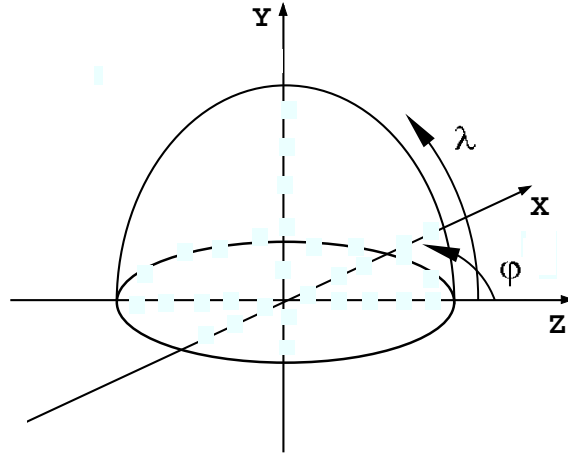


Figure 6.3: Longitude and Latitude Angles. φ and λ are measured for all views as depicted in this schema.

(φ_3, λ_3) the north view. (For the case of reconstruction from two views only condition (iv) has to be fulfilled.) The following equations are given for only one object point, the position of which is to be determined for the unfamiliar view. The coordinates of this point in the sample views are given by (x_i, y_i) for $i = 1, 2, 3$. Its coordinates (\hat{x}, \hat{y}) in the unfamiliar view are to be found.

Two Sample Views

x-coordinate: If

(b1 i) φ_1 and φ_2 are not 180° apart and

(b1 ii) φ_1 and φ_2 are not equal,

the x -coordinate \hat{x} is a linear combination of x_1 and x_2 . The coefficients of this linear combination are simple functions in φ_1, φ_2 , and $\hat{\varphi}$. In detail:

$$\hat{x} = \sum_{i=1}^2 a_i x_i \quad (6.1)$$

with

$$a_1 = -\csc(\varphi_1 - \varphi_2) \cdot \sin(\varphi_2 - \hat{\varphi}), \quad (6.2)$$

$$a_2 = \csc(\varphi_1 - \varphi_2) \cdot \sin(\varphi_1 - \hat{\varphi}). \quad (6.3)$$

y-coordinate: The y -coordinate \hat{y} is a linear combination of x_1, x_2 , and y_1 if, again, (b1 i) and (b1 ii) are fulfilled and

(b2iii) the center view is not positioned in the north pole of the viewing hemisphere.

The coefficients of this linear combination are more complex. Two of them depend on $\varphi_1, \varphi_2, \hat{\varphi}, \lambda_1$, and $\hat{\lambda}$. The third of them depends on λ_1 and $\hat{\lambda}$ only. In detail:

$$\hat{y} = \sum_{i=1}^2 b_i \cdot x_i + b_3 \cdot y_1 \quad (6.4)$$

with

$$b_1 = \cos(\varphi_2 - \hat{\varphi}) \cdot \csc(\varphi_1 - \varphi_2) \cdot \sin(\hat{\lambda}) - \cos(\hat{\lambda}) \cdot \cot(\varphi_1 - \varphi_2) \cdot \tan(\lambda_1), \quad (6.5)$$

$$b_2 = \csc(\varphi_1 - \varphi_2) \cdot \left(\cos(\hat{\lambda}) \cdot \tan(\lambda_1) - \cos(\varphi_1 - \hat{\varphi}) \cdot \sin(\hat{\lambda}) \right), \quad (6.6)$$

$$b_3 = \cos(\hat{\lambda}) \cdot \sec(\lambda_1). \quad (6.7)$$

Three Sample Views

x-coordinate: For three sample views the linear combination for the x -coordinate \hat{x} is the same as for two sample views. It depends only on x_1 and x_2 if the same conditions (**b1 i**) and (**b1 ii**) as for two sample views hold.

y-coordinate: The y -coordinate \hat{y} is a linear combination of y_1, y_2 , and y_3 if (**b1 i**), (**b1 ii**),

(**b4iii**) λ_1 and λ_3 are not equal and

(**b4 iv**) the center view is not positioned on the equator of the viewing hemisphere.

The coefficients depend on $\varphi_1, \varphi_2, \hat{\varphi}, \lambda_1, \lambda_3$, and $\hat{\lambda}$. In detail:

$$\hat{y} = \sum_{i=1}^3 b_i y_i \quad (6.8)$$

with

$$\begin{aligned} b_1 = & \csc(\lambda_1 - \lambda_3) \cdot \csc(\varphi_1 - \varphi_2) \cdot \\ & \left[\cos(\hat{\varphi}) \cdot \sin(\hat{\lambda}) \cdot \left(\cos(\lambda_3) \cdot \sin(\varphi_2) - \cot(\lambda_1) \cdot \sin(\lambda_3) \cdot \sin(\varphi_1) \right) + \right. \\ & \left. \sin(\hat{\lambda}) \cdot \sin(\hat{\varphi}) \cdot \left(\cos(\lambda_3) \cdot \cos(\varphi_2) - \cos(\varphi_1) \cdot \cot(\lambda_1) \cdot \sin(\lambda_3) \right) - \right. \\ & \left. \cos(\hat{\lambda}) \cdot \sin(\lambda_3) \cdot \sin(\varphi_1 - \varphi_2) \right], \end{aligned} \quad (6.9)$$

$$b_2 = \csc(\lambda_1) \cdot \csc(\varphi_2 - \varphi_1) \cdot \sin(\hat{\lambda}) \cdot \sin(\varphi_1 + \hat{\varphi}), \quad (6.10)$$

$$\begin{aligned} b_3 = & \csc(\lambda_1 - \lambda_3) \cdot \csc(\varphi_1 - \varphi_2) \cdot \\ & \left(\cos(\hat{\lambda}) \cdot \sin(\lambda_1) \cdot \sin(\varphi_1 - \varphi_2) + \right. \\ & \left. \cos(\lambda_1) \cdot \sin(\hat{\lambda}) \cdot \left(\sin(\varphi_1 + \hat{\varphi}) - \sin(\varphi_2 + \hat{\varphi}) \right) \right). \end{aligned} \quad (6.11)$$

The calculation of only one point position (\hat{x}, \hat{y}) of the set of object point positions for an unfamiliar view has been described now. These calculations are performed for all points. After rounding to integer values this results in the set $\hat{\mathcal{X}}_{(\hat{p}, \hat{q})}$ of object point positions for the unfamiliar view $(\hat{p}, \hat{q}) = (\hat{\varphi}, \hat{\lambda})$.

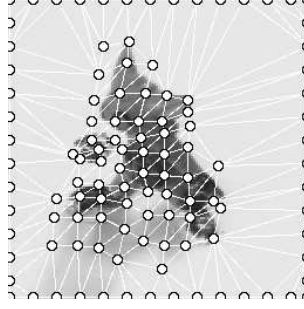


Figure 6.4: Triangulation. The Delaunay triangulation of the set $\hat{\mathcal{X}}_{(\hat{p}, \hat{q})} \cup \hat{\mathcal{Y}}_{(\hat{p}, \hat{q})}$ is superimposed on the original image of view $(\hat{p}, \hat{q}) = (7, 11)$ of the “Tom” object, which is to be morphed (compare with figure 6.6).

6.1.2 Warping From Familiar to Unfamiliar Views

In this subsection the second step of the view morphing algorithm is described. Given the original gray level image $I_{(p,q)}$ of the center view (p, q) of the view bubble in question together with its original object point positions $\mathcal{X}_{(p,q)}$ and given the set $\hat{\mathcal{X}}_{(\hat{p}, \hat{q})}$ of object point positions for the unfamiliar view (\hat{p}, \hat{q}) the task is now to derive the final morphed image $\hat{I}_{(\hat{p}, \hat{q})}$ of the unfamiliar view (\hat{p}, \hat{q}) . This morphed image is obtained by *warping* image $I_{(p,q)}$ according to the transformation which is provided by $\mathcal{X}_{(p,q)}$ and $\hat{\mathcal{X}}_{(\hat{p}, \hat{q})}$. Wolberg [83] defines the term *digital image warping* as “geometric transformation of digital images”. It means the redefinition of the spatial relationship between points in an image.

To obtain $\hat{I}_{(\hat{p}, \hat{q})}$, sets $\mathcal{Y}_{(p,q)}$ and $\hat{\mathcal{Y}}_{(\hat{p}, \hat{q})}$ of auxiliary vertices are inserted at the boundaries of $I_{(p,q)}$ and $\hat{I}_{(\hat{p}, \hat{q})}$, respectively. The subjoined vertices are positioned for both images at the same locations, that is to say, equidistantly and starting at position $(0, 0)$ in accordance with the generation of grid graphs described in section 3.5. Then the point set $\hat{\mathcal{X}}_{(\hat{p}, \hat{q})} \cup \hat{\mathcal{Y}}_{(\hat{p}, \hat{q})}$ is triangulated by applying an algorithm proposed by Mehlhorn and Näher [37], which starts with an arbitrary triangulation and afterwards performs Delaunay flips to obtain a Delaunay triangulation (see figure 6.4). The triangulation of $\hat{\mathcal{X}}_{(\hat{p}, \hat{q})} \cup \hat{\mathcal{Y}}_{(\hat{p}, \hat{q})}$ is conferred to the corresponding vertices in the set $\mathcal{X}_{(p,q)} \cup \mathcal{Y}_{(p,q)}$. Both triangulations then allow a simple, linear interpolation of pixel positions inside corresponding triangles yielding the sought morphed view $\hat{I}_{(\hat{p}, \hat{q})}$. This last step is described now.

Triangle Interpolation for View Morphing

To each pixel of image \hat{I} the gray value of the corresponding pixel in image I should be assigned. Each pixel of \hat{I} lies in one triangle provided by the preceding Delaunay triangulation of $\hat{\mathcal{X}} \cup \hat{\mathcal{Y}}$. The following derivation is described for one triangle only (see figure 6.5).

For each $\hat{\mathbf{v}}$ in a triangle $(\hat{\mathbf{s}}, \hat{\mathbf{t}}, \hat{\mathbf{u}})$ in \hat{I} the corresponding position \mathbf{v} in the corresponding triangle $(\mathbf{s}, \mathbf{t}, \mathbf{u})$ in I is to be located with $\hat{\mathbf{s}}, \hat{\mathbf{t}}, \hat{\mathbf{u}} \in \hat{\mathcal{X}} \cup \hat{\mathcal{Y}}$, and $\mathbf{s}, \mathbf{t}, \mathbf{u} \in \mathcal{X} \cup \mathcal{Y}$. To this end $\hat{\mathbf{v}} = \begin{pmatrix} \hat{v}_1 \\ \hat{v}_2 \\ 1 \end{pmatrix}$ can be expressed as linear combination of $\hat{\mathbf{s}}, \hat{\mathbf{t}}$, and $\hat{\mathbf{u}}$ with coefficients c_1, c_2, c_3 with $\sum_{i=1}^3 c_i = 1$, which means that c_1, c_2 , and c_3 are the *areal coordinates* of $\hat{\mathbf{v}}$ with

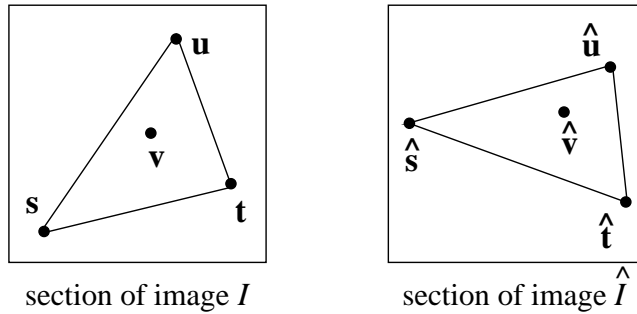


Figure 6.5: Corresponding triangles in I and \hat{I} . See text for a description.

respect to the triangle $(\hat{s}, \hat{t}, \hat{u})$:

$$\hat{\mathbf{v}} = \widehat{M} \cdot \mathbf{c} \quad (6.12)$$

with

$$\widehat{M} = \begin{pmatrix} \hat{s}_1 & \hat{t}_1 & \hat{u}_1 \\ \hat{s}_2 & \hat{t}_2 & \hat{u}_2 \\ 1 & 1 & 1 \end{pmatrix}. \quad (6.13)$$

$\det \widehat{M} \neq 0$ is met for the non-degenerated triangles (as they are provided by the applied triangulation algorithm). This allows the notation

$$\mathbf{c} = \widehat{M}^{-1} \cdot \hat{\mathbf{v}}. \quad (6.14)$$

In the same manner one can express the sought vector \mathbf{v} as linear combination of \mathbf{s} , \mathbf{t} , and \mathbf{u} with the *same* coefficients c_i :

$$\mathbf{v} = M \cdot \mathbf{c} \quad (6.15)$$

with

$$M = \begin{pmatrix} s_1 & t_1 & u_1 \\ s_2 & t_2 & u_2 \\ 1 & 1 & 1 \end{pmatrix}. \quad (6.16)$$

Along with equation 6.14 this provides \mathbf{v} by

$$\mathbf{v} = M \cdot \widehat{M}^{-1} \cdot \hat{\mathbf{v}}. \quad (6.17)$$

Finally, after rounding \mathbf{v} to integer values, the pixel at position $\hat{\mathbf{v}}$ in \hat{I} gets the gray level value of the pixel at position \mathbf{v} in image I . As the whole image is partitioned into triangles this procedure yields the morphed image \hat{I} . Figures 6.6 and 6.7 give two examples of morphed views.

6.2 Evaluation of Morphed Views

In the previous section was described how an arbitrary unfamiliar view of an object can be generated from two or three sample views of the object. Now I will turn to the question of the quality of such a morphed view. For that purpose a relative error between a morphed view and its referring original view is calculated. This is described in subsection 6.2.1. In

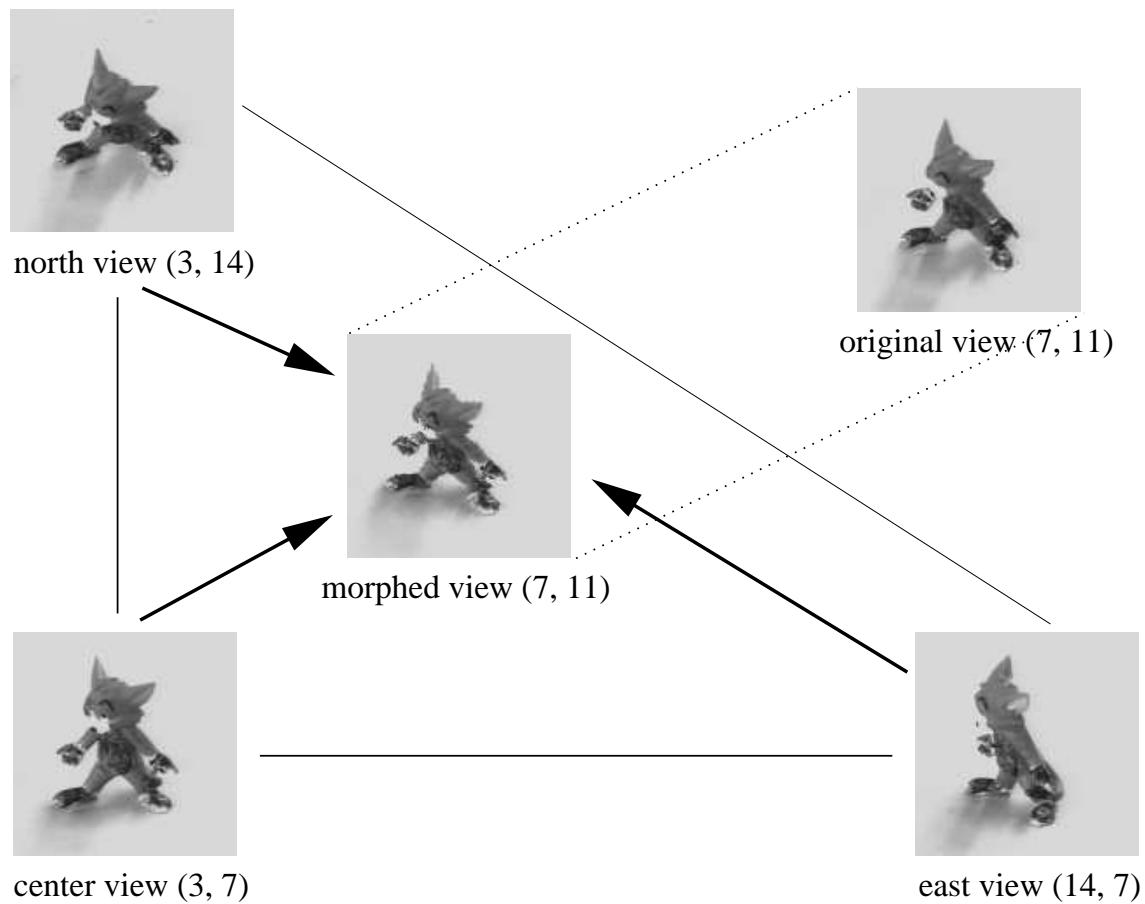


Figure 6.6: Example of Morphed View for Object “Tom”. View (7, 11) is morphed from the three sample views (3, 7), (14, 7), and (3, 14).

the remaining subsections the dependency of the relative errors of morphed views on the numbers of view bubbles which constitute the object representation and on the tracking threshold which is used to generate the view bubbles is examined by a statistical description of large data sets of errors obtained experimentally.

6.2.1 Relative Errors

Usually in error calculation one is interested in the *relative error* between the true value of a quantity (in this case $I_{(\hat{p}, \hat{q})}$) and a measured or inferred value (in this case $\hat{I}_{(\hat{p}, \hat{q})}$). The relative error is defined by the difference between the true and the measured value divided by the true value without taking the absolute value of this quotient [78]. However, often only the absolute value of the relative error is of interest. For one pixel of the images in question this absolute relative error is defined by $|(i - \hat{i})/i|$ if i and \hat{i} are the gray values of corresponding pixels in I and \hat{I} . As this definition comprises a disadvantageous dependency of the error on the intensity of the concerned pixel, the error is calculated not

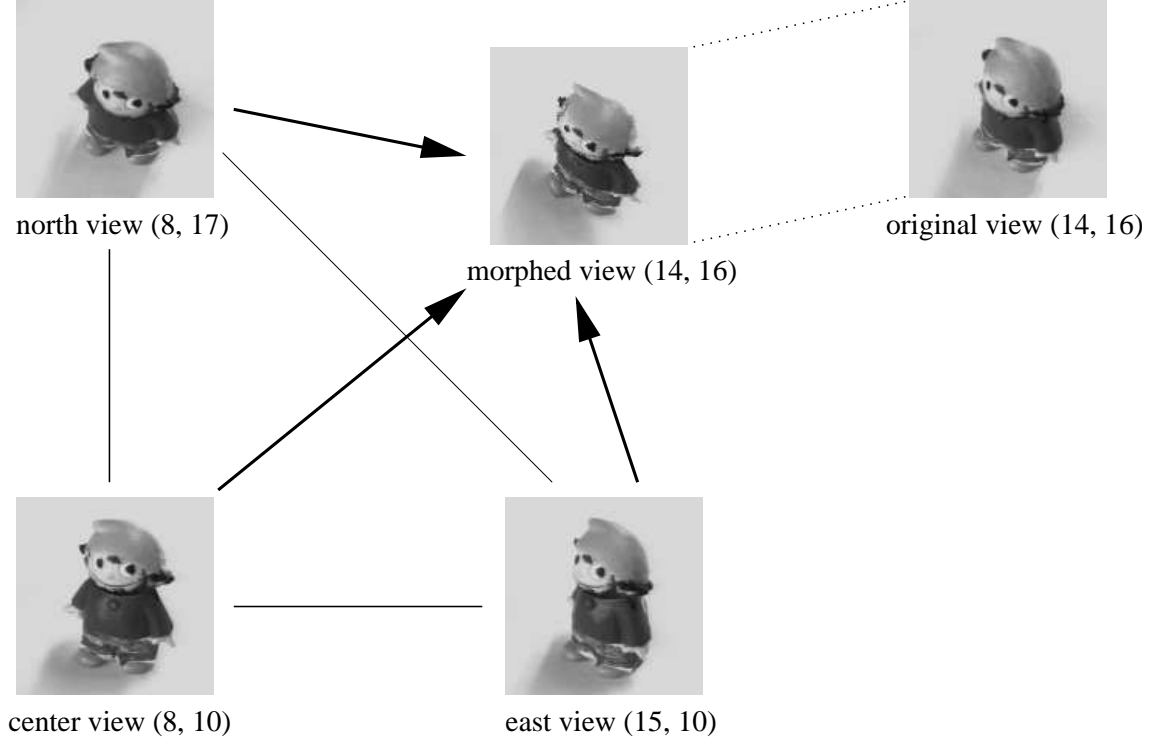


Figure 6.7: Example of Morphed View for Object “Dwarf”. View (14, 16) is morphed from the three sample views (8, 10), (15, 10), and (8, 17).

relative to the true value i but relative to the range of values of the images:

$$e_{max}(I, \hat{I}) = \max_{i \in I, \hat{i} \in \hat{I}} |i - \hat{i}| \quad (6.18)$$

where i and \hat{i} are gray values of arbitrary pixels in I and \hat{I} , rather than gray values of corresponding pixels. Now a (preliminary) relative error between I and \hat{I} can be defined by

$$\epsilon'_{morph}(I, \hat{I}) := \frac{1}{N \cdot M} \cdot \frac{1}{e_{max}(I, \hat{I})} \cdot \sum_{j=1}^{N \cdot M} |i_j - \hat{i}_j| \quad (6.19)$$

where $N \cdot M$ is the size of the images. To be robust against slight translations of the object in the image plane \hat{I} is shifted across I with a small offset (13 pixels) in all four directions. For all shifting positions ϵ'_{morph} is calculated. Then the final relative error between \hat{I} and I is defined as the minimal ϵ'_{morph} over all shifting positions:

$$\epsilon_{morph}(I, \hat{I}) := \min_{\text{shift pos.}} \epsilon'_{morph}(I, \hat{I}). \quad (6.20)$$

6.2.2 Methods

In the last chapter the sparse representation \mathcal{R} of an object has been introduced, which depends on the similarity threshold τ of the tracking algorithm and examples of \mathcal{R} for

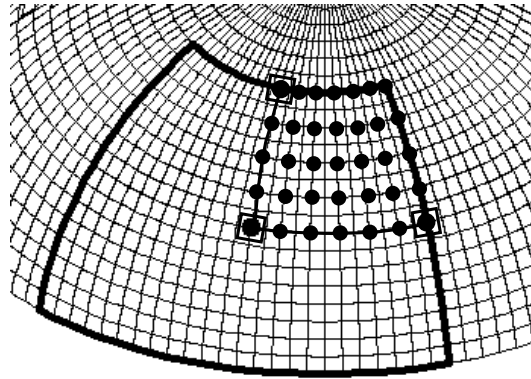


Figure 6.8: Positions of Unfamiliar Views Which are Morphed for Statistics. For each view bubble of a particular partitioning of the viewing hemisphere, which is given by a sparse representation \mathcal{R} , a morphed version of the unfamiliar views marked by dots is generated. As described in section 6.1 in the case of three sample views the object point positions of each unfamiliar view are calculated from the point positions in the center, the north and the east border view of the view bubble, which are marked by small squares (compare with figure 6.2) whereas in the case of two sample views only the center and east views are used.

five different values of τ have been given in section 5.2. For each of these partitionings of the viewing hemisphere unfamiliar views are morphed now. For each view bubble of a particular partitioning views are morphed in the first quadrant as depicted in figure 6.8. Only every second view is generated, and that from two and from three sample views as described in section 6.1.¹ Then for each morphed view \hat{I} the relative error $\epsilon_{morph}(I, \hat{I})$ is calculated. This provides ten data sets for each object: “relative errors of morphed views generated from two sample views” and “relative errors of morphed views generated from three sample views” for five different partitionings each. For each data set the mean and the maximal relative errors are calculated and depicted in diagrams depending on the number ν of view bubbles in \mathcal{R} and on the similarity threshold τ for tracking. In addition, the Levenberg-Marquardt method for non-linear parameter fitting [40] is performed again to fit the mean errors depending on ν . The fittings are performed separately for both objects as well as separately for morphing from two or three sample views, respectively. The resulting parameters for the fitting curves of these four data sets are then averaged and the final curve is displayed in the diagrams with the mean and maximal relative errors. The same procedure is performed for fitting the mean errors depending on τ . This allows a comparison of the quality of morphed views between the tested objects and between morphing from two and from three sample views.

¹For the linear combination of object point positions the conditions (b1 i) to (b4 iv) must be met. This is guaranteed by excluding some view bubbles from the statistic evaluation. In detail: (b1 i) and (b1 ii) are met for all view bubbles under inspection, whereas (b2 iii), (b4 iii), or (b4 iv) are violated for the “Tom” object twice for $\tau = 0.85$, once for $\tau = 0.9$, and 89 times for $\tau = 0.95$, and for the “dwarf” object once for $\tau = 0.85$, once for $\tau = 0.9$, and 44 times for $\tau = 0.95$.

6.2.3 Results

First let me make some remarks on ϵ_{morph} . The use of only one single parameter to measure the quality of a morphed view holds the danger that very different artefacts can yield the same value of the parameter. The results for ϵ_{morph} show an increasing quality of morphed views for a decreasing relative error. That means, the visual impression correlates with the performance of ϵ_{morph} . Furthermore, the relative errors grow larger if the distance between the unfamiliar view (\hat{p}, \hat{q}) and the sample view (p, q) is increasing. Both effects are demonstrated in figure 6.9 and support the reasonability of using ϵ_{morph} as a measure of quality.

Figure 6.10 shows the mean and maximal relative errors and the fitting curves depending on the number of view bubbles in the representations of the “Tom” object and the “dwarf” object, respectively. The mean relative errors for both objects for two as well as three sample views are monotonically decreasing with an increasing number of view bubbles in the object representation \mathcal{R} , i.e., with a decreasing size of the view bubbles. The correlation between the mean errors and the number ν of view bubbles can be expressed by a logarithmic function

$$g(\nu) = c \cdot \ln \nu + d \quad (6.21)$$

for $\nu \in [4, 289]$, which is the range of numbers of view bubbles. Figure 6.11 shows the same data plotted against the similarity threshold τ for the tracking algorithm which was used to generate the object representation. Here the fitting function

$$h(\tau) = a \cdot \tau + b, \quad (6.22)$$

$\tau \in [0.75, 0.95]$, for the mean errors is linearly decreasing. These approximations are consistent with the exponential fitting function $\nu = f(\tau)$ described in section 5.2 in the last chapter (see figure 5.5 and equation 5.2) and they allow the estimation of the necessary number of view bubbles in \mathcal{R} (or the necessary similarity threshold for the generation of the representation) for a given mean relative error of the morphed unfamiliar views. For example, if the mean relative error should be 3% equation 6.21 provides a value of about 73 view bubbles necessary in the sparse representation.

The figures 6.10 and 6.11 reveal smaller relative errors for the “Tom” object than for the “dwarf” object.

Surprisingly at a first glance, the errors for morphing from two sample views are only slightly larger than the errors for morphing from three sample views; for the finest partitioning of the viewing hemisphere they are even smaller than for three sample views.

6.2.4 Discussion

The morphing experiments have been performed for every second view in the first quadrant of each view bubble only. I assume that the results can be transferred to the remaining views, which have not been tested. Concerning the quality of the morphed views two weaknesses can be noticed. Both are inherent in the problem of view synthesis from sample views. The first is connected with self-occlusions of object parts. Only parts of the object which are visible in image $I_{(p,q)}$ can be mapped in image $\hat{I}_{(\hat{p},\hat{q})}$. A demonstration of this effect is shown in figure 6.9. View (13, 13) is morphed incorrectly, because the tail of “Tom”, which is visible in the original view (13, 13) is not visible in the source view (3, 7).

The second weakness can be explained by a visibility constraint called *monotonicity*, as Seitz and Dyer [64] point out. For two sample views this constraint requires that all visible object points appear in the same order in the sample views. Then, theoretically, intermediate views can be synthesized unambiguously, otherwise unfamiliar views cannot be predicted exactly. For three sample views the pairwise monotonicity between every pair of views inside the triangle spanned by the sample views permits the synthesis of any view in this triangle. Anyway, the range of views that can be predicted may not be larger than a single aspect of an aspect graph (see section 2.1.2). As monotonicity in general cannot be measured a priori I can only assume that for smaller view bubbles, i.e., for larger similarity thresholds τ , it is met, whereas for larger view bubbles it cannot be guaranteed. Both of these general problems can account for the fact that the quality of the morphed views is decreasing with a decreasing number of view bubbles in \mathcal{R} as well as with an increasing distance of the unfamiliar view from the sample views. In theory there are no restrictions on the position of the unfamiliar view in relation to the sample views (see section 6.1.1 where no conditions on $\hat{\varphi}$ and $\hat{\lambda}$ are established). Thus, a large distance should provide equal quality morphs as small distances as long as the monotonicity constraint is met and no self-occlusions occur. Of course, the results also depend on the density and the positions $\mathcal{X}_{(p,q)}$ of the graph vertices in the source view $I_{(p,q)}$ and on the quality of the correspondences provided by the tracking algorithm, i.e., the point sets $\mathcal{X}_{(p,q)}^e$ and $\mathcal{X}_{(p,q)}^n$.

One possible reason why the relative errors for the “Tom” object are in general smaller than for the “dwarf” object may lie in the fact that self-occlusions occur earlier for the “dwarf” because of its convex shape (see figure 3.2 for an example). These self-occlusions can lead to more faulty reconstructions.

The very small differences measured between the relative errors for morphing from two and from three sample views are not really surprising taking into account, that from the formulas given in section 6.1.1 no difference at all is expected for ideal conditions, i.e., for compliance of monotonicity and absence of self-occlusions within the range of reconstruction. Thus, the differences which are measured can only be reducible to deviations from the ideal conditions. And, in fact, for larger view bubbles in the representation the difference between two and three sample views is more distinctive than for smaller view bubbles (particularly obvious in figure 6.11 for the “dwarf” object). This is a hint that the differences occur due to self-occlusions or a violated monotonicity constraint, which are more likely for larger distances between the sample views. In other words, for a large distance between two sample views a third sample view can provide more information for a better reconstruction than for a smaller distance between two sample views. Furthermore, the differences between morphing from two and from three sample views are more distinctive for the “dwarf” object, which can be explained by the same argument: more self-occlusions occur for the “dwarf” than for “Tom”. For very small distances between sample views, as given for the finest partitioning of the viewing hemisphere, almost no self-occlusions or violations of the monotonicity constraint should occur. In this case the same results from two as well as three sample views could be expected. However, the data which represents the third view is erroneous due to the imperfect tracking of corresponding object points. Thus, for very small distances between the sample views the third view probably adds an additional error instead of compensating for the error introduced by the second view. This can explain the fact, that for the finest partitioning of the viewing hemisphere two sample views provide even better results than three sample views.

Another interesting aspect of view synthesis from three sample views which could have been investigated is the question of differences in error rates depending on the position of the synthesized view inside or outside of the triangle spanned by the sample views. This is implied by the triangles in the first row of figure 6.9. The effect of the unfamiliar view's position inside or outside the triangle has not been analyzed in detail, but the fact that there are gradually increasing relative errors instead of an abrupt rise when the triangle is left suggests that the inside/outside condition does not have a strong effect on the results.

I would like to make some remarks on the biological plausibility of object perception by a linear combination of sample views. As mentioned earlier in this chapter (section 6.1.1) Ullman and Basri's approach [74] is of theoretical importance and has not been applied to real-world images, because perfect correspondences between the sample views have to be provided and no self-occlusions of the objects are allowed to occur. The psychophysical study carried out by Bülthoff and Edelman [9] mentioned in section 2.2.4 yielded a rejection of this linear combination model. But they also base their study upon the assumption of perfect correspondences and very few self-occlusions. For example, in parts they use wire-like test objects. If one starts from realistic assumptions such as real-world objects with self-occlusions and imperfect correspondences as done for this thesis, my results yielded by the linear combination approach are consistent with Bülthoff and Edelman's findings. For instance, the decreasing recognition rates for views which lie outside the connecting line between two sample views have been a reason for them to reject the linear combination model. But this is exactly what my linear combination approach under realistic assumptions provides: decreasing quality of an unfamiliar view with an increasing distance from the sample views.

Summarizing, I can say that a mean relative error of morphed views of about 5% for a reasonable partitioning of the viewing hemisphere seems to be small enough not to be able to question the *linear combination* of object point positions in familiar views to generalize to unfamiliar views. This partly answers question **Q4** from the introduction (chapter 1).

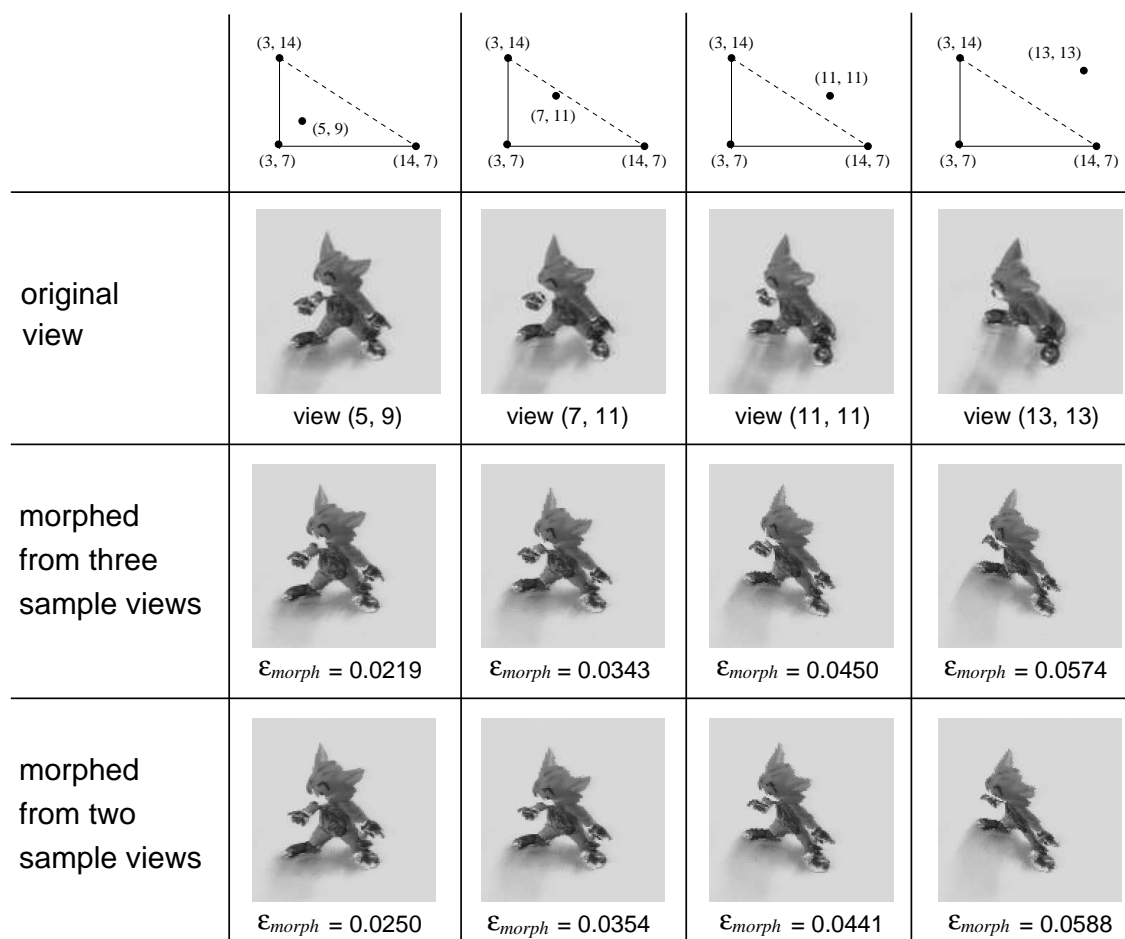


Figure 6.9: Performance of the Relative Error. The last two rows of this diagram show unfamiliar views of the “Tom” object which were morphed from two or three sample views. For each case the sample views were the same as depicted in figure 6.6. Only the position of the unfamiliar view varies from column to column. The first row shows the positions of the views in relation to each other. In the second row the original images of the unfamiliar views are displayed to be compared with their morphed versions. For two as well as for three sample views the relative error increases with an increasing distance of the unfamiliar view from view (3, 7), which is the source of the gray level values for the morphed view. The quality of the morphed views assessed by visual inspection decreases with an increasing distance as well. The differences between morphing from two and from three sample views are negligible for this example.

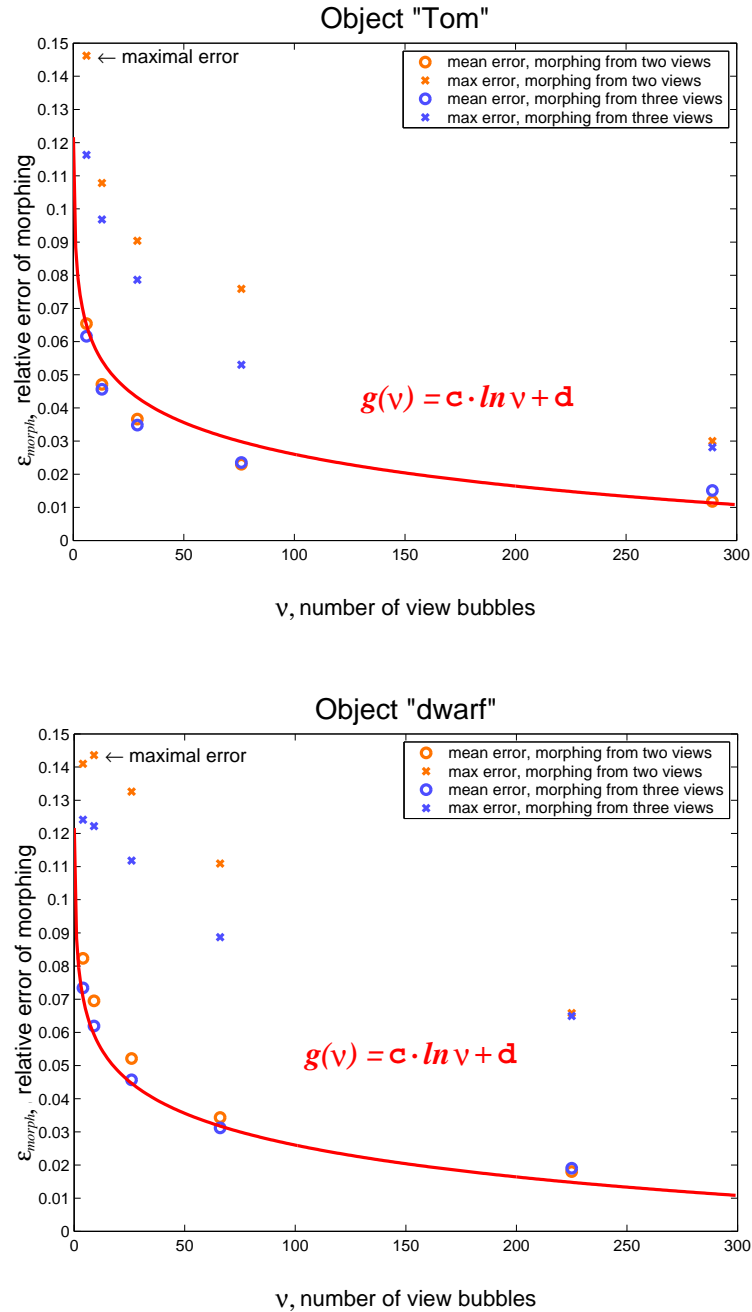


Figure 6.10: Correlation Between Morphing Errors and Number of View Bubbles. The fitting function is calculated from the data sets for both objects and has the parameters $c = -0.014$ and $d = 0.09$.

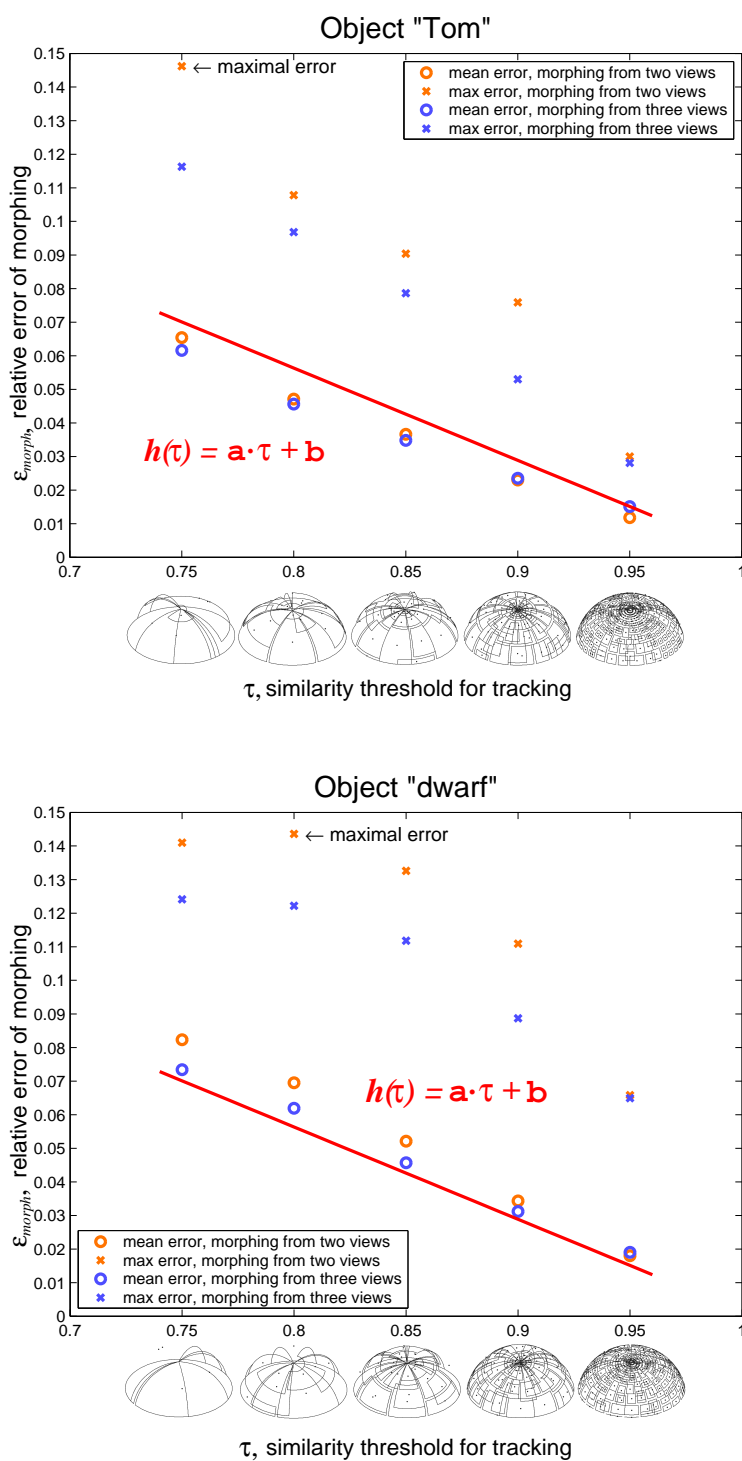


Figure 6.11: Correlation between Morphing Errors and Similarity Threshold. The fitting function is calculated from the data sets for both objects and has the parameters $a = -0.275$ and $b = 0.276$.

Chapter 7

Virtual Views

In the last chapter the possibility to generate *images* of unfamiliar views of an object from the few views stored in its sparse representation \mathcal{R} was demonstrated. However, as the main goal of this thesis is the modelling of perception functions such as the estimation of an object’s pose, rather than view synthesis, it is necessary to generate *representations* of unfamiliar views from views constituting \mathcal{R} . In equation 3.9 the representation of a view has been introduced in terms of pairs of vertex positions and feature vectors. In the last chapter of these two only the vertex positions of unfamiliar views from sample views were calculated. In this chapter, now, the calculation also of feature vectors from sample views will be described, resulting in complete representations of unfamiliar views, as needed for perception. The feature vectors of unfamiliar views are derived by an *interpolation* of feature vectors in sample views, introduced in subsection 7.1.1. A resulting view representation can be visualized by a reconstruction of the view from the new feature vectors. The reconstructed view is denoted by the term “virtual view”. This is described in subsection 7.1.2. The quality of the representations of unfamiliar views is assessed in this chapter in section 7.2 by an error analysis and a statistical description of a large set of virtual views, similar to the evaluation of morphed views in the last chapter. Furthermore, the quality of the generated representations of unfamiliar views has to be proved in chapter 8, where they are applied to pose estimation.

7.1 Virtual View Generation

The same processing scheme as the one used for the generation of morphed views described in section 6.1 is applied here to generate virtual views. Virtual views are generated exclusively inside view bubbles because only there correspondences are available. The sample views for the generation of virtual views are the same as in chapter 6 in the case of two as well as three sample views. They depend on the virtual view’s position inside its view bubble.

The generation of virtual views proceeds in three steps (see figure 7.1):

1. The first step is the same as for the generation of morphed views: object point *positions* of the unfamiliar view are derived by *linear combinations* of the corresponding point positions in the sample views, resulting in the set $\hat{\mathcal{X}}_{(\hat{p}, \hat{q})}$ of point positions in the unfamiliar view (\hat{p}, \hat{q}) , as described in subsection 6.1.1.

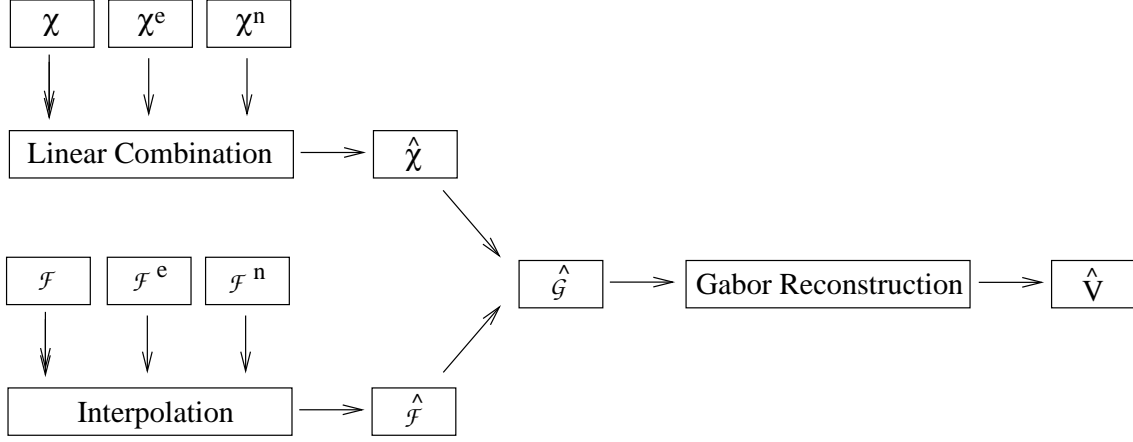


Figure 7.1: Flowchart of the Generation of Virtual Views From Three Sample Views. A virtual view of an object is generated in three steps. Firstly, the positions $\mathcal{X}, \mathcal{X}^e$ and \mathcal{X}^n of object points in the sample views are linearly combined to the unfamiliar view's set of point positions $\hat{\mathcal{X}}$ as described in subsection 6.1.1 (see equations 3.9 and 4.1). Secondly, an interpolation between the feature vectors $\mathcal{F}, \mathcal{F}^e$ and \mathcal{F}^n at object points in the sample views is performed yielding the set $\hat{\mathcal{F}}$ of feature vectors, which describes the corresponding object points in the unfamiliar view. The first and second step provide the sought representation $\hat{\mathcal{G}}$ of an unfamiliar view. From this calculated representation the virtual view \hat{V} is reconstructed in the third step.

2. In the second step the object point *features* of the unfamiliar view are derived by *interpolations* between features at corresponding object points in the sample views. For the example of three source views with the unfamiliar view lying in the first quadrant of its view bubble the problem can be formulated as follows:

given: the sets $\mathcal{F}_{(p,q)}, \mathcal{F}_{(p,q)}^e$ and $\mathcal{F}_{(p,q)}^n$ of object point features in the sample views and the positions of the sample views and the unfamiliar view on the viewing hemisphere,

sought: the set $\hat{\mathcal{F}}_{(\hat{p},\hat{q})}$ of object point features in an unfamiliar view (\hat{p}, \hat{q}) which lies in the first quadrant of view bubble $B_{(p,q)}$.

For the case of other quadrants or only two sample views the given feature sets vary as described in section 6.1.

In subsection 7.1.1 a solution for this problem is proposed.

After this second step the sought representation $\hat{\mathcal{G}}_{(\hat{p},\hat{q})}$ of an unfamiliar view (\hat{p}, \hat{q}) is available, because the sets $\hat{\mathcal{X}}_{(\hat{p},\hat{q})}$ and $\hat{\mathcal{F}}_{(\hat{p},\hat{q})}$ have only to be combined to

$$\hat{\mathcal{G}}_{(\hat{p},\hat{q})} = \langle \hat{\mathcal{X}}_{(\hat{p},\hat{q})}, \hat{\mathcal{F}}_{(\hat{p},\hat{q})} \rangle \quad (7.1)$$

(compare with equation 3.9).

3. In the third step a virtual view is generated by an *image reconstruction from the interpolated Gabor wavelet responses* of graph $\hat{\mathcal{G}}_{(\hat{p},\hat{q})}$. This is described in more detail in subsection 7.1.2.

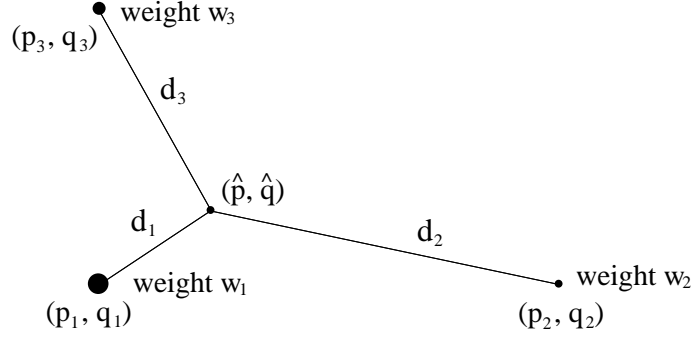


Figure 7.2: Weighting of Feature Vectors for Three Sample Views. In this example, w_1 is the strongest weight, because the unfamiliar view (\hat{p}, \hat{q}) is closer to sample view (p_1, q_1) than to the sample views (p_2, q_2) and (p_3, q_3) .

7.1.1 Interpolation of Object Point Features

In this subsection formulas are given for step two of the algorithm for generating virtual views. Let $\hat{\mathcal{J}}_{(\hat{p}, \hat{q})}$ denote one feature vector of the set $\hat{\mathcal{F}}_{(\hat{p}, \hat{q})}$ of all feature vectors of the unfamiliar view (\hat{p}, \hat{q}) (compare with section 3.5). $\hat{\mathcal{J}}_{(\hat{p}, \hat{q})}$ should be interpolated from the corresponding feature vectors $\mathcal{J}_{(p_1, q_1)}$, $\mathcal{J}_{(p_2, q_2)}$, and $\mathcal{J}_{(p_3, q_3)}$ in the sample views (p_1, q_1) , (p_2, q_2) , and (p_3, q_3) (see figure 7.2). The order of numeration is the same as in subsection 6.1.1. The interpolation consists in calculating a weighted mean. In the case of two sample views, $\hat{\mathcal{J}}_{(\hat{p}, \hat{q})}$ is calculated as weighted sum of $\mathcal{J}_{(p_1, q_1)}$ and $\mathcal{J}_{(p_2, q_2)}$, in the case of three sample views it is a weighted sum of $\mathcal{J}_{(p_1, q_1)}$, $\mathcal{J}_{(p_2, q_2)}$ and $\mathcal{J}_{(p_3, q_3)}$:

$$\hat{\mathcal{J}}_{(\hat{p}, \hat{q})} = \sum_{i=1}^S w_i \mathcal{J}_{(p_i, q_i)} \quad (7.2)$$

where $S = 2$ or 3 is the number of sample views and $\sum_{i=1}^S w_i = 1$. The weights w_i depend on the relative position of the unfamiliar view with respect to the sample views. A smaller distance between the unfamiliar view and a sample view leads to a stronger weight than a larger distance. Let $d_i := d\left(\begin{pmatrix} \hat{p} \\ \hat{q} \end{pmatrix}, \begin{pmatrix} p_i \\ q_i \end{pmatrix}\right)$ be the Euclidean metric for $i = 1, \dots, S$. Then the weights are calculated depending on the number of used sample views as listed in the next two brief sections.

Two Sample Views

$$w_1 = d_2 / (d_1 + d_2), \quad (7.3)$$

$$w_2 = d_1 / (d_1 + d_2). \quad (7.4)$$

Three Sample Views

$$w_1 = d_2 d_3 / (d_1 d_2 + d_1 d_3 + d_2 d_3), \quad (7.5)$$

$$w_2 = d_1 d_3 / (d_1 d_2 + d_1 d_3 + d_2 d_3), \quad (7.6)$$

$$w_3 = d_1 d_2 / (d_1 d_2 + d_1 d_3 + d_2 d_3). \quad (7.7)$$

At this point the complete representation $\widehat{\mathcal{G}}_{(\hat{p}, \hat{q})}$ of an unfamiliar view, which can be utilized for perception tasks, has been derived from the representations $\mathcal{G}_{(p_i, q_i)}$ of the sample views, $i = 1, \dots, S$.

7.1.2 Virtual View Reconstruction

In this subsection the reconstruction of images from labeled graphs in general, proposed by Pötzsch et al. [57, 58], is introduced and the third step, that of the generation of virtual views, is explained. In the last subsection the interpolation of feature vectors has been described. Feature vectors throughout this thesis are Gabor filter responses, which have been introduced in subsection 3.4.2. The Gabor transform was described by the operator \mathcal{W} :

$$\mathcal{J}_{\vec{k}} = (\mathcal{W}I)_{\vec{k}}$$

for a Gabor kernel specified by the parameter \vec{k} (see equation 3.6). Because of the linearity of the operator \mathcal{W} the reconstructed version \tilde{I} of I should be retrievable from the values $\mathcal{J}_{\vec{k}}$ by a linear combination

$$\tilde{I}(\vec{x}) = \mathcal{V}\tilde{\mathcal{J}} = \sum_{\vec{k}} \mathcal{J}_{\vec{k}} b_{\vec{k}}(\vec{x}) \quad (7.8)$$

with appropriate basis functions $b_{\vec{k}}(\vec{x})$. Since the Gabor filters $\psi_{\vec{k}}$ are not orthonormal it is necessary to choose a linear combination of $\overline{\psi}_{\vec{k}}$ for the basis functions $b_{\vec{k}}$ instead of the functions $\overline{\psi}_{\vec{k}}$ themselves:

$$b_{\vec{k}}(\vec{x}) = \sum_{\vec{l}} (\Psi^{-1})_{\vec{k}\vec{l}} \overline{\psi}_{\vec{l}}(\vec{x}) \quad \text{with} \quad \Psi_{\vec{k}\vec{l}} := \int \overline{\psi}_{\vec{k}}(\vec{x}) \psi_{\vec{l}}(\vec{x}) d\vec{x}. \quad (7.9)$$

The latter dot products of the Gabor filters can be computed analytically, which is done in [57].

As a single object view is represented by a graph labeled with jets (which are vectors of Gabor wavelet responses) it is interesting to visualize the amount of information in such a representation by a reconstruction of the view from the Gabor responses. For this purpose, the reconstruction described above can be approximated by a *local* reconstruction of each jet restricted to a Voronoi area around its location. This provides the image V reconstructed from its original graph \mathcal{G} as displayed in figure 7.3.

As the feature vectors of the sample views, from which the feature vectors of the unfamiliar view have been interpolated, are Gabor wavelet responses, also the *interpolated* vectors can be interpreted as Gabor wavelet responses. A reconstruction $\widehat{V}_{(\hat{p}, \hat{q})}$ from these “virtual” jets of the representation $\widehat{\mathcal{G}}_{(\hat{p}, \hat{q})}$ of an unfamiliar view (\hat{p}, \hat{q}) is performed and completes the last step of the generation of virtual views. An example of a reconstructed image $\widehat{V}_{(\hat{p}, \hat{q})}$ is displayed in figure 7.3 as well.

7.2 Evaluation of Virtual Views

In the last section the generation of the representation $\widehat{\mathcal{G}}$ of an unfamiliar view was derived. It has been shown that this representation can be visualized by a reconstruction of a virtual view \widehat{V} from $\widehat{\mathcal{G}}$. This virtual view can be helpful exploring the quality of the

representation $\widehat{\mathcal{G}}$ by a comparison with an image V which is reconstructed from the original graph \mathcal{G} of the unfamiliar view using the same reconstruction method. In this comparison the images \widehat{V} and V appear in the same roles as the morphed image \widehat{I} and its original version I in section 6.2 in the last chapter. The evaluation of the quality of the virtual views has been reported by Peters and von der Malsburg [52] and is performed in analogy to the evaluation of morphed views described in section 6.2. In particular, the same relative error is calculated between V and \widehat{V} , which is denoted by $\epsilon_{virt}(V, \widehat{V})$ in this chapter (compare with formula 6.20), and analogous statistics are evaluated for large data sets of errors obtained experimentally.

7.2.1 Methods

Actually, it is sufficient to refer to subsection 6.2.2 to describe the methods used to analyze the quality of virtual views, and thus the quality of generated representations of unfamiliar views. Each detail of the methods used for assessing morphed views is transferred to the methods of evaluating virtual views. In particular, for each of the five different partitionings of a viewing hemisphere, which have been introduced in chapter 5, virtual views are generated according to the scheme depicted in figure 6.8: in the case of two sample views the object point features of an unfamiliar view marked by a dot are interpolated from the object point features in the center and east view of the view bubble, in the case of three sample views they are calculated from the center, the east, and the north view. The interpolation of feature vectors functions as described in subsection 7.1.1. After the generation of a virtual view \widehat{V} and the image V for each of these unfamiliar views and after calculating $\epsilon_{virt}(V, \widehat{V})$, the data sets “relative errors of virtual views generated from two sample views” and “relative errors of virtual views generated from three sample views” are obtained for each partitioning of the hemisphere for both objects. With these data sets the same parameter fitting is performed as described in subsection 6.2.2 for morphed views. The resulting data and curves for virtual views are depicted in diagrams to compare the quality of virtual views between both objects and between an interpolation of feature vectors from two and from three sample views.

7.2.2 Results

The results obtained in this chapter resemble those obtained in chapter 6 and are depicted in figures 7.4 and 7.5. Figure 7.4 shows the mean and maximal relative errors and the fitting curves depending on the number of view bubbles in the representations of the “Tom” object and the “dwarf” object, respectively. As for the mean relative errors of morphed views the mean relative errors of virtual views are monotonously decreasing with an increasing number of view bubbles in the object representation \mathcal{R} for both objects and for two as well as three sample views. The correlation between the mean errors and the number ν of view bubbles can be expressed by the qualitatively same logarithmic function 6.21 with different parameters.

Figure 7.5 shows the same data plotted against the similarity threshold τ for the tracking algorithm which was used to generate the object representation. Here the fitting function is also a linearly decreasing function 6.22, also with slightly different parameters as in chapter 6. The fitting functions allow the estimation of the necessary number of

view bubbles in \mathcal{R} (or the necessary similarity threshold for the generation of the representation) for a given mean relative error of the virtual views and thus for a given quality of representations of unfamiliar views. For example, if the mean relative error should be 3% equation 6.21 provides a value of about 166 view bubbles necessary in the object representation.

As in the case of morphed views the figures 7.4 and 7.5 reveal smaller relative errors for the “Tom” object than for the “dwarf” object. In addition, the errors for the feature interpolation from two sample views are again only slightly larger than for the interpolation from three sample views and for the finest partitioning of the viewing hemisphere they are even smaller than for three sample views, which is another parallel to the results of the last chapter.

7.2.3 Discussion

Each of the results described in the last subsection can be explained by the same phenomena as described in detail in subsection 6.2.4. The decreasing error rates for an increasing number of view bubbles in the object representation, as well as the lower error rates for the “Tom” object than for the “dwarf” object, as well as the slight advantage of an interpolation from three sample views compared to an interpolation from two sample views for larger view bubbles, as well as the reversed relationship for the finest partitioning of the viewing hemisphere can be reduced to self-occlusions and the monotonicity constraint.

As on the one hand, object recognition with graphs labeled with Gabor wavelet responses has been proven to be very powerful (for example by Lades et al. [29]) and as on the other hand, the errors of the reconstruction of an unfamiliar view from its inferred representation $\hat{\mathcal{G}}$ are quite small (for instance, about 5% for a convenient partitioning of the viewing hemisphere), I expect reasonable object recognition rates also from $\hat{\mathcal{G}}$, but of course this still has to be proved by experiments. To verify the applicability of $\hat{\mathcal{G}}$ for perception tasks pose estimation experiments have been performed, which are described in the next chapter. Another approach to obtain the representation $\hat{\mathcal{G}}$ of an unfamiliar view could be the extraction of $\hat{\mathcal{G}}$ from the morphed version \hat{I} of the unfamiliar view. In this thesis this was not subjected to investigation.

Concluding I can say that, together with the results of chapter 6, the results of this chapter allow an answer to question **Q4** posed in the introduction. The linear combination of object point positions and the interpolation of object point features can be assessed as useful strategies to combine familiar to unfamiliar views. Because of the almost perfect reconstruction results for morphed as well as virtual views from sample views with small distances I suppose that the boundaries of the range of generalization from familiar to unfamiliar views is determined only by specific object properties such as self-occlusions instead of limitations of the combining methods introduced in the last chapters.

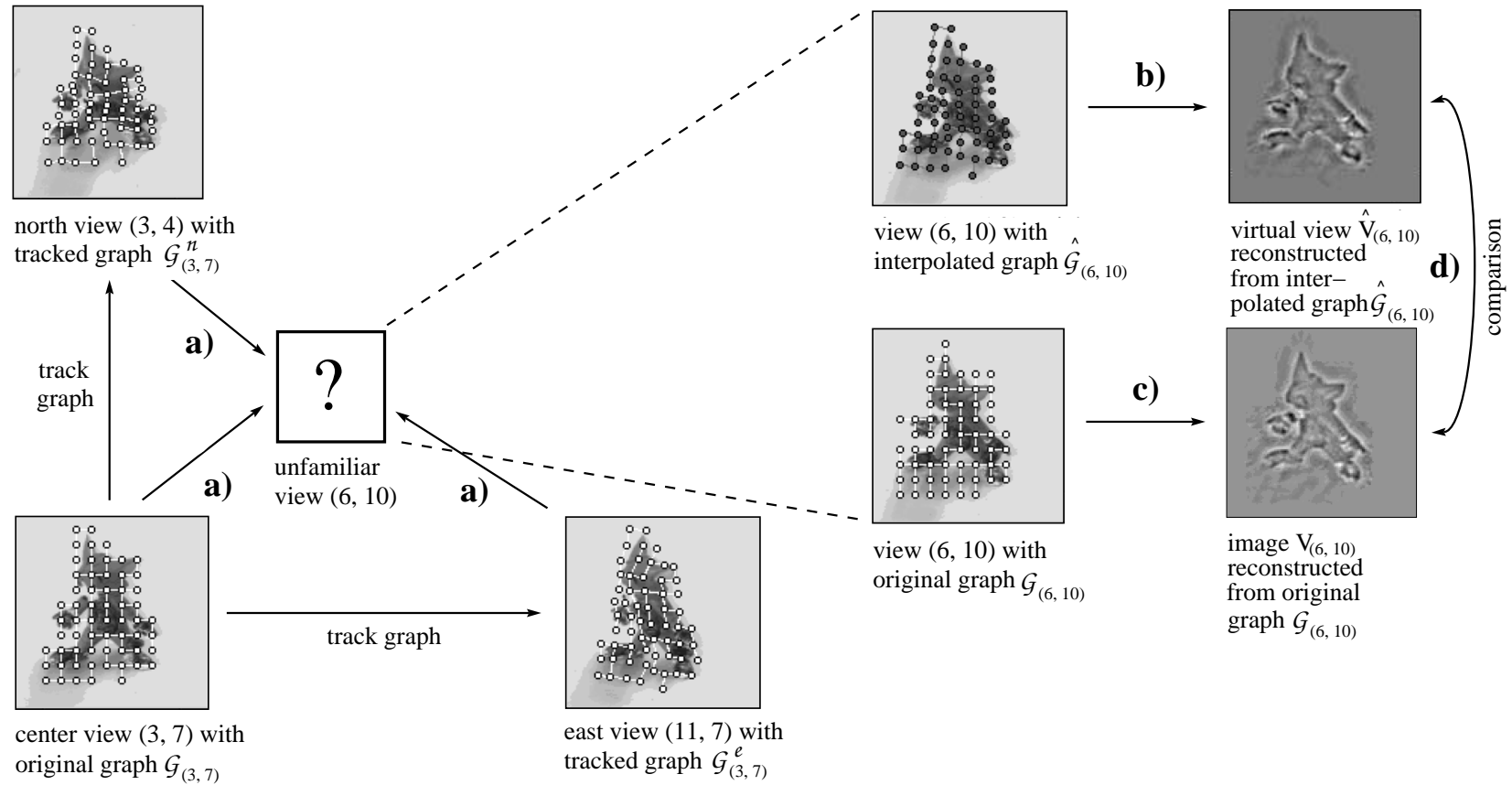


Figure 7.3: Example of Virtual View Generation. **a)** In the first two steps of the algorithm the interpolated graph $\hat{\mathcal{G}}_{(\hat{p},\hat{q})} = \hat{\mathcal{G}}_{(6,10)}$, which represents the unfamiliar view $(\hat{p},\hat{q}) = (6,10)$, is calculated by a linear combination of vertex positions and an interpolation of feature vectors of the sample views (center, east, and north view). **b)** In the third step of the algorithm the virtual view $\hat{V}_{(\hat{p},\hat{q})} = \hat{V}_{(6,10)}$ is reconstructed from $\hat{\mathcal{G}}_{(6,10)}$. **c)** View (6, 10) can also be reconstructed from its original graph $\mathcal{G}_{(\hat{p},\hat{q})} = \mathcal{G}_{(6,10)}$ yielding image $V_{(\hat{p},\hat{q})} = V_{(6,10)}$. **d)** To evaluate the quality of the information in the interpolated graph $\hat{\mathcal{G}}_{(6,10)}$, $V_{(6,10)}$ and $\hat{V}_{(6,10)}$ can be compared.

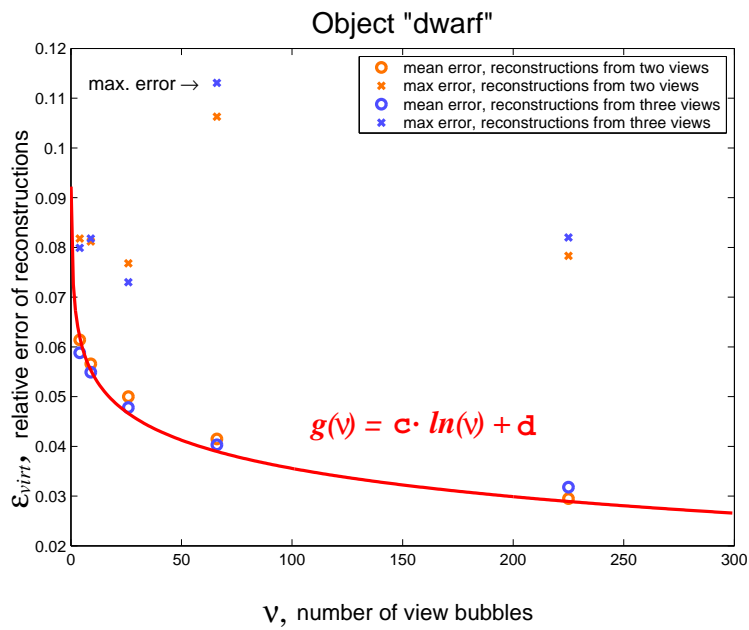
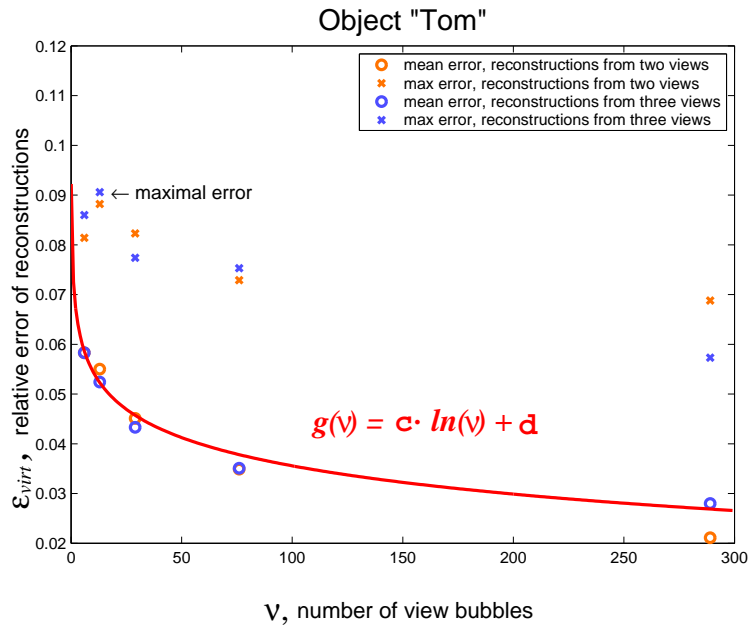


Figure 7.4: Correlation Between Virtual View Errors and Number of View Bubbles. The fitting function is calculated from the data sets for both objects and has the parameters $c = -0.009$ and $d = 0.076$.

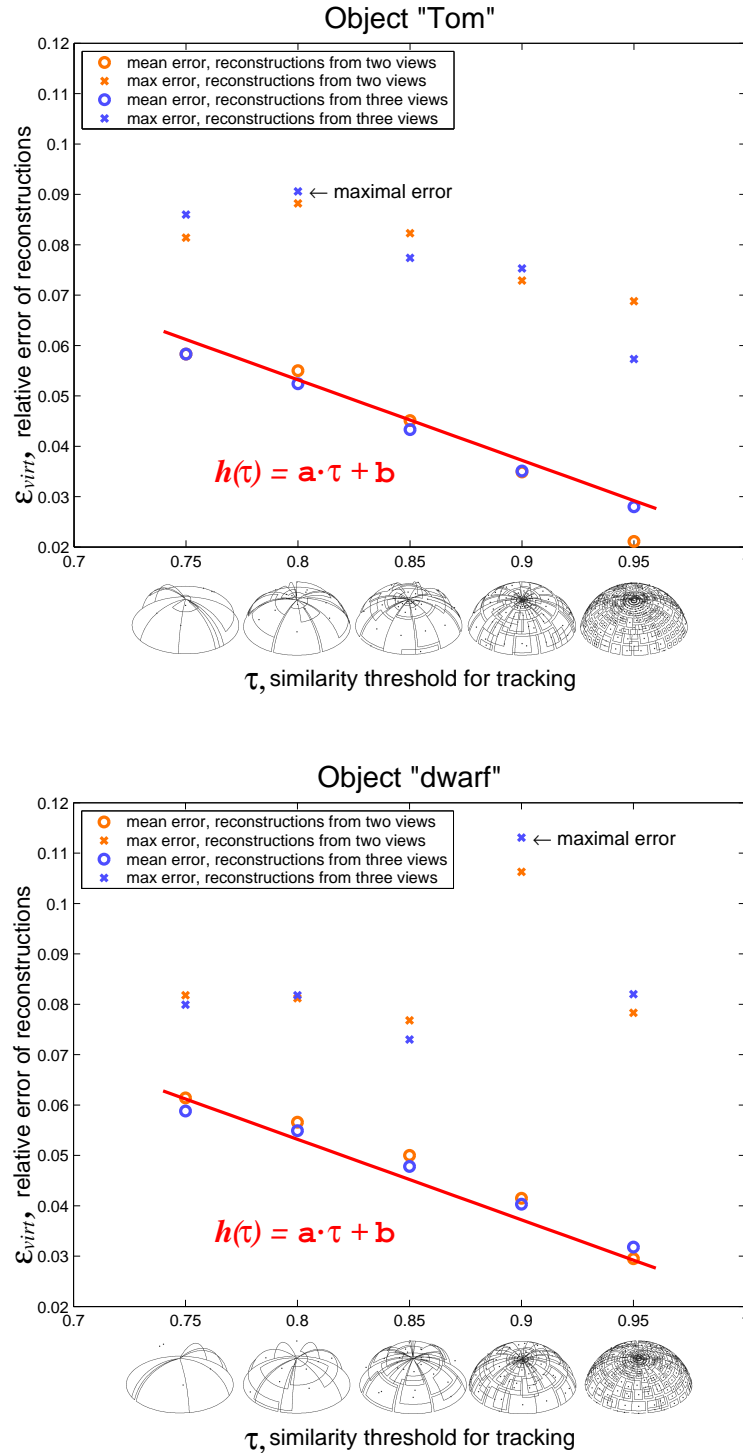


Figure 7.5: Correlation between Virtual View Errors and Similarity Threshold. The fitting function is calculated from the data sets for both objects and has the parameters $a = -0.156$ and $b = 0.179$.

Chapter 8

Pose and Sequence Estimation

In the last chapters the sparse representation \mathcal{R} of a three-dimensional object has been derived, which consists of representations of only some views of the object. Furthermore, it has been shown that the representations of views not contained in \mathcal{R} can be derived from the sample views. Herewith the information necessary to perform perception tasks is available. In this chapter the concept of view-based object perception is verified by using \mathcal{R} and the view generating techniques to estimate the pose of an object, which is an essential function of perception, e.g., when the object is to be grasped. The poses of smooth test sequences of the rotating objects are estimated from the original (i.e., non-degraded) images and from images with additive noise. Two different estimation experiments are performed. On the one hand, each view of a test sequence is treated independently. These experiments are denoted by the term “single pose estimation”. On the other hand, the neighborhood of successive views of a test sequence can be exploited. Experiments which take this temporal context into account are denoted by the term “sequence estimation”. In the first two sections of this chapter experiments with non-degraded images are described. Section 8.1 deals with single pose estimation, section 8.2 with sequence estimation. In section 8.3 single pose as well as sequence estimation experiments with test sequences with a substantial amount of noise are described.

8.1 Single Pose Estimation

Given the sparse representation of the object in question and given a test view of the object, the aim is the determination of the object’s pose displayed in the test view, i.e., the assignment of the test view to its correct position on the viewing hemisphere. In this section a solution to this problem is proposed (subsection 8.1.1) and the results of simulations with a series of test views are reported (subsection 8.1.2) and discussed (subsection 8.1.3).

8.1.1 Methods

Let T be the test view, the pose of which should be estimated, and \mathcal{G}_T be its representing graph, which is extracted from the original image of view T as described in section 3.5 after the test view has been divided into object and background segments as described in section 3.3. This means that no a priori knowledge about the object is provided. Remember that a view is determined by its position on the viewing hemisphere (see section 3.2)

and remember the sparse representation $\mathcal{R} = \{\mathcal{B}_i\}_{i \in R} = \{\langle \mathcal{G}_i, \mathcal{G}_i^w, \mathcal{G}_i^e, \mathcal{G}_i^s, \mathcal{G}_i^n \rangle\}_{i \in R}$ of an object where R is the cover of the viewing hemisphere defined in subsection 5.1.1. Let $I_i, i \in R$, be the center images of the view bubbles the graphs \mathcal{G}_i are extracted from. The *single pose estimation algorithm* for estimating the pose of a single test view T proceeds in two steps:

1. Match \mathcal{G}_T to each image $I_i, i \in R$, as described in section 3.6 using the similarity function \mathcal{S}_{abs} (equation 3.7). As a *rough estimate* of the object's pose choose that view bubble \hat{B} the center image I_i of which provides the largest similarity to \mathcal{G}_T .
2. Generate the representation $\hat{\mathcal{G}}$ for each unfamiliar view which is included in \hat{B} using the techniques described in subsections 6.1.1 and 7.1.1 for two sample views. From each of the calculated graphs $\hat{\mathcal{G}}$ generate the corresponding virtual view \hat{V} (see subsection 7.1.2). Compare each of the virtual views \hat{V} with the virtual view \hat{V}_T reconstructed from \mathcal{G}_T using the error function $\epsilon_{virt}(\hat{V}, \hat{V}_T)$ (see section 7.2). The estimated pose \hat{T} of the test view T is the position on the viewing hemisphere of that virtual view \hat{V} which provides the smallest error ϵ_{virt} .

The *estimation error* between T and \hat{T} can be determined by the Euclidean distance as it has already been used in subsection 7.1.1:

$$\epsilon_{esti}(T, \hat{T}) = d(T, \hat{T}). \quad (8.1)$$

For the reconstruction of the virtual views only two sample views are used, because of the insignificant differences between reconstructions from two and three sample views (see figures 7.4 and 7.5 in subsection 7.2.2).

For the evaluation of the algorithm three sequences \mathcal{T} of ten test views each have been chosen, which are used throughout this chapter. They are displayed in figure 8.1. For both objects and for each partitioning of the viewing hemisphere the poses of these 30 test views have been estimated.

8.1.2 Results

The illustrations in figure 8.1 indicate that pose estimation becomes more precise with an increasing number of sample views in the object representation. This result has been expected and is confirmed by an inspection of the mean estimation errors taken over the 30 test views for each object and each partitioning of the hemisphere separately. They are summarized in the first column of table 8.1. With one exception for the ‘‘Tom’’ object the mean errors are decreasing with an increasing value of τ . Figure 8.2 shows an example of estimated poses.

8.1.3 Discussion

The results of the single pose estimation experiments for non-degraded images are amazingly good. This is particularly obvious for the example displayed in figure 8.2, taking into account that the sparse representation of the ‘‘Tom’’ object contains only the representations of those 30 views depicted in figure 5.6. This was the test sequence for which the best result for $\tau = 0.75$ was obtained, but also for a reasonable partitioning of the

single pose estimation, non-degraded images

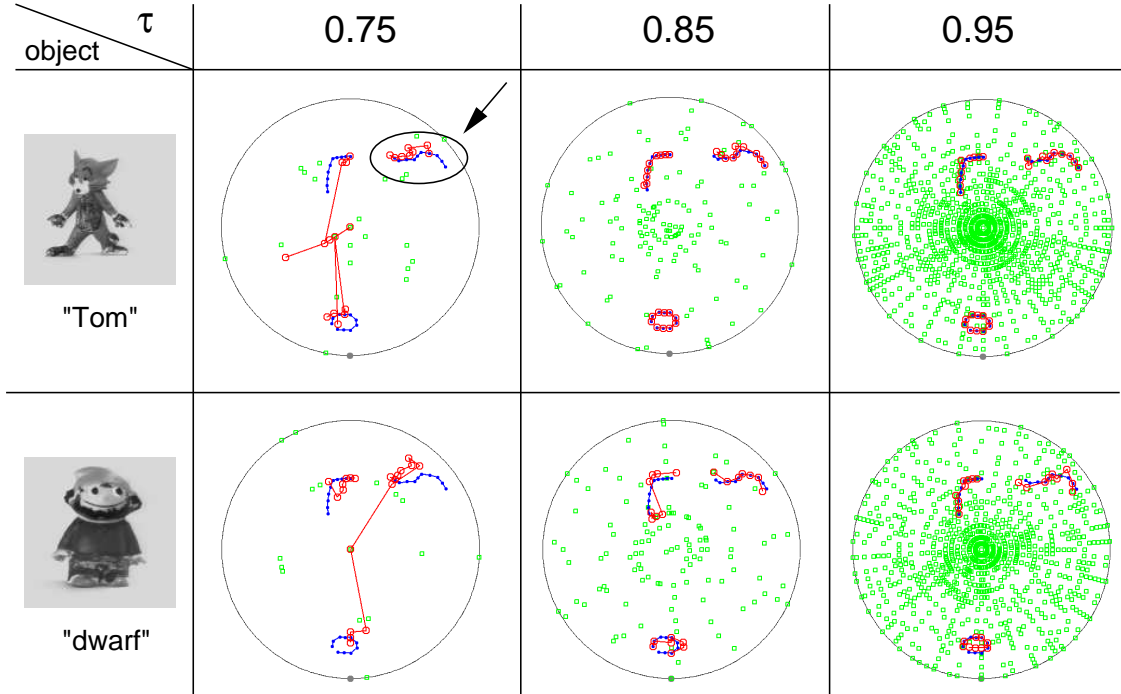


Figure 8.1: Single Pose Estimation for Non-Degraded Images. For three partitionings of the viewing hemisphere the results of the single pose estimation experiments are depicted. The light gray squares indicate the views which are represented in the object representation \mathcal{R} , black dots mark the positions of the three test sequences \mathcal{T} , and the estimated positions are tagged by dark gray circles. The arrow points at the test sequence and its estimation which is displayed in figure 8.2.

viewing hemisphere ($\tau = 0.85$) the mean estimation errors are smaller than 5° for both objects, which can be regarded as a remarkable result. This provides another argument which supports the good quality of the calculated representations $\hat{\mathcal{G}}$ of the unfamiliar views and allows the conclusion that the proposed view-based approach to object perception is suitable for pose estimation.

mean errors, non-degraded images			
	τ	single pose estimation	sequence estimation
object "Tom"	0.75	36.51°	33.91°
	0.8	3.63°	3.49°
	0.85	0.77°	0.77°
	0.9	3.35°	1.76°
	0.95	0.36°	0.12°
object "dwarf"	0.75	20.54°	9.75°
	0.8	19.47°	3.38°
	0.85	4.2°	2.9°
	0.9	2.65°	1.36°
	0.95	1.71°	0.77°

Table 8.1: Mean Estimation Errors for Non-Degraded Images. Each value given in degrees is the mean estimation error computed from the 30 test views. For example, for the "Tom" object and the partitioning of $\tau = 0.75$ the average distance of the estimated pose \hat{T} to the true pose T is 36.51° if each view has been estimated independently (first column). In the second column the mean errors are presented for the sequence estimation experiments, which are described in section 8.2. For the "dwarf" object $mean_{single} > mean_{sequence}$ is significant for $\tau = 0.75$ on a 10%-level, for $\tau = 0.8$ on a 5%-level, for $\tau = 0.85$ on a 10%-level, for $\tau = 0.9$ on a 1%-level, and for $\tau = 0.95$ on a 2.5%-level, which has been ascertained with the one-tailed t -test. The significant values are enclosed in a rectangle.

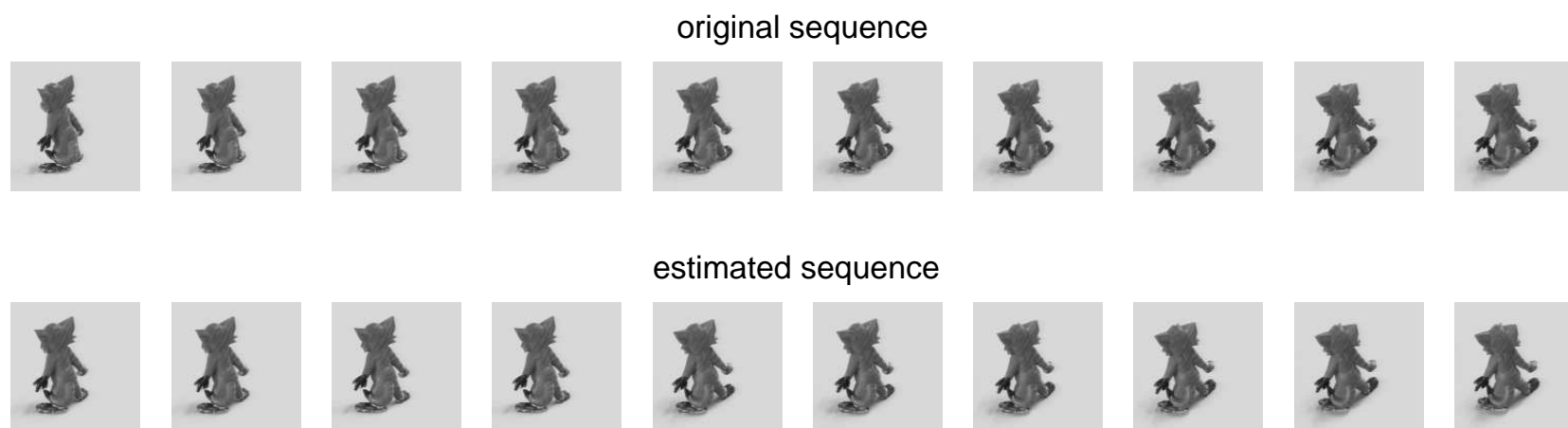
single pose estimation, object "Tom", $\tau=0.75$ 

Figure 8.2: Example for Single Pose Estimation. This figure shows the images of the test sequence and its estimation which are marked in figure 8.1. For this example the representation of the “Tom” object for $\tau = 0.75$ has been chosen. It consists of the views which are depicted in figure 5.6. In the first row the true poses of the object, which should be estimated, are displayed. The second row shows the poses which have been estimated by treating each view of the sequence independently. The estimation error for this sequence averages 5.78° .

8.2 Sequence Estimation

Although the results reported in the last section are already satisfactory it could be possible to achieve an improvement of the pose estimation results by incorporating information which has not been utilized up to now. The temporal context of object views, which are presented to an observer while the object rotates, could provide additional, useful hints to the object's pose. In subsection 8.2.1 I propose a pose estimation procedure which takes into account that successive views of the test sequences are arranged contiguously on the viewing hemisphere. Experiments are performed with the same test sequences as in the previous section, the results of which are reported in subsection 8.2.2 and discussed in subsection 8.2.3.

8.2.1 Methods

The views of the test sequences already used for single pose estimation are neighbors on the viewing hemisphere, which means that the x - as well as the y -coordinate of the positions of successive views deviate at most by one unit. Furthermore, no view appears more than once in a sequence.

Let $\hat{n} = |\hat{B}|$ be the number of stored views contained in the view bubble \hat{B} which is provided by step one of the single pose estimation algorithm for a single test view, and let \hat{N} be the smaller of \hat{n} and 10: $\hat{N} = \min(\hat{n}, 10)$. The results of the single pose estimation can be improved by a *sequence estimation algorithm* which consists of two mechanisms. The first mechanism can provide an improvement by fine-tuning inside the rough estimate \hat{B} , the second mechanism can correct single outliers of the estimated sequence if the rough estimate failed.

1. **“Fine-tuning” mechanism:** For each view of the test sequence \mathcal{T} the first step of the single pose estimation algorithm is performed. Then, step two of the single pose estimation algorithm is extended by the determination of a *series* of best estimates of each view of the test sequence, instead of determining only one best estimate. In detail, let $T_j, j = 1, \dots, 10$, be the views of the test sequence \mathcal{T} . The \hat{N}_j best estimates $\hat{T}_{jk}, k = 1, \dots, \hat{N}_j$, of test view T_j are the positions on the viewing hemisphere of those virtual views which provide the \hat{N}_j smallest errors ϵ_{esti} (compare with subsection 8.1.1). This implies that they are contained in the same view bubble, which has been determined by step one of the single pose estimation algorithm. The values \hat{T}_{jk} are numbered in descending order of similarity, i.e., \hat{T}_{j1} is the best estimate of view T_j . In contrast to the single pose estimation, where \hat{T}_{j1} is the estimated pose of T_j , here the estimated pose is chosen from the set of the \hat{N}_j best estimates, taking into account that \hat{T}_{jk} should be neighboring to its predecessor $\hat{T}_{j-1,k'}$.

The fine-tuning mechanism starts with the addition of \hat{T}_{11} to an empty, preliminary estimated sequence \hat{T}_{pre} . If $\hat{T}_{j-1,k'}$ has been added to \hat{T}_{pre} and \hat{T}_{jk} is sought as an estimate of T_j , a loop runs for $k = 1, \dots, \hat{N}_j$ and checks the neighborhood, i.e., the angular distance, of $\hat{T}_{j-1,k'}$ and \hat{T}_{jk} . The first \hat{T}_{jk} which is neighboring to $\hat{T}_{j-1,k'}$ is added to \hat{T}_{pre} , but only if it has not been added as estimate of an earlier view of \mathcal{T} . If this loop is left without having found any estimate of T_j , the first \hat{T}_{jk} is added to \hat{T}_{pre} which has not been added before regardless of its neighborhood to $\hat{T}_{j-1,k'}$. Only

if this search is not successful either, \widehat{T}_{j1} is chosen as estimate of T_j . This procedure provides the preliminary estimated sequence \widehat{T}_{pre} .

As each of the \widehat{N}_j best estimates of a view T_j is contained in the same view bubble, which has been determined as rough estimate of view T_j , this fine-tuning mechanism can usually provide an improvement of the single pose estimation result only if the rough estimation of T_j has provided the correct view bubble.

2. **“Outlier” mechanism:** As the first step of the single pose estimation algorithm does not always provide satisfactory results, it is possible that $\widehat{T}_{\text{pre}} := (\widehat{T}_1, \dots, \widehat{T}_{10})$ still contains some outliers where the rough pose estimation of a single view failed and the fine-tuning mechanism could not yield an improvement. Single outliers can be corrected by the following mechanism. An estimated view $\widehat{T}_j, j = 2, \dots, 9$, which has a large angular distance (in the sense of equation 8.1) to its predecessor \widehat{T}_{j-1} as well as to its successor \widehat{T}_{j+1} , is corrected by choosing the mean position instead: $\widehat{T}_j := \frac{1}{2}(\widehat{T}_{j-1} + \widehat{T}_{j+1})$. As a limit of distance 8 units on the viewing hemisphere have been chosen.

\widehat{T}_1 and \widehat{T}_{10} are treated separately thereafter. If $d(\widehat{T}_1, \widehat{T}_2) \geq 8$ than \widehat{T}_1 is replaced by $2\widehat{T}_2 - \widehat{T}_3$. This means that an outlier at the beginning of \widehat{T}_{pre} is replaced by a view which provides a smooth transition to the second and third estimated views of the sequence, because \widehat{T}_2 takes now the mean position between the new pose \widehat{T}_1 and \widehat{T}_3 . (In case that the substitute for \widehat{T}_1 does not lie in the range of the viewing hemisphere, \widehat{T}_1 is replaced by that view inside the range of the hemisphere which lies closest to $2\widehat{T}_2 - \widehat{T}_3$.) For an outlying pose \widehat{T}_{10} at the end of \widehat{T}_{pre} the corresponding procedure is performed. This results in the final estimated sequence \widehat{T} .

To evaluate the quality of the sequence estimation algorithm the poses of the test sequences \mathcal{T} have been estimated for both objects and each partitioning of the viewing hemisphere.

8.2.2 Results

In the second column of table 8.1 the mean estimation errors computed from the three test sequences with 10 views each are summarized for the sequence estimation. As in the case of single pose estimation the mean errors are decreasing with an increasing value of τ with one exception for the “Tom” object.

As expected, a comparison of the mean estimation error $mean_{\text{single}}$ of the single pose estimation with the mean estimation error $mean_{\text{sequence}}$ of the sequence estimation yields better results for the sequence estimation algorithm. For all partitionings and for both objects $mean_{\text{sequence}} \leq mean_{\text{single}}$ holds. For the “dwarf” object $mean_{\text{single}} > mean_{\text{sequence}}$ is significant for all partitionings of the hemisphere, which has been ascertained with the one-tailed t -test .

Most of the improvements achieved by the sequence estimation are due to the fine-tuning mechanism, because for most of the views to be estimated the first step of the single pose estimation already provides the correct view bubble. This effect is exemplified in figure 8.3, where three sequences and their typical estimates are shown for a reasonable partitioning of the viewing hemisphere.

8.2.3 Discussion

The figures 7.4 and 7.5 in the last chapter indicated a better quality of the calculated representations $\hat{\mathcal{G}}$ of unfamiliar views for the “Tom” object than for the “dwarf” object. This may be a reason why the difference between the results from single pose and sequence estimation is significant for the “dwarf” object, whereas for the “Tom” object it is not. The better quality of the generated representations of unfamiliar views implies a better result of the single pose estimation for the “Tom” object than for the “dwarf” object. That means that the aid of the additional neighborhood information provided by the fine-tuning mechanism of the sequence estimation algorithm has a larger effect on the improvement of the results for the “dwarf” object than for the “Tom” object. In general, the results reported in the last subsection allow for the conclusion that the integration of information from successive views can improve the estimate of an object’s pose from non-degraded images.

non-degraded images


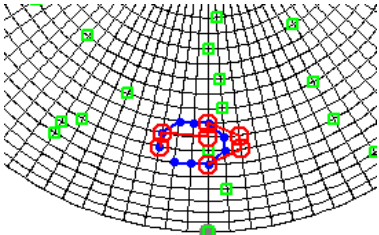
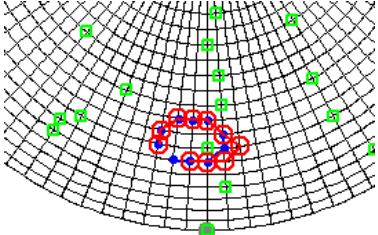

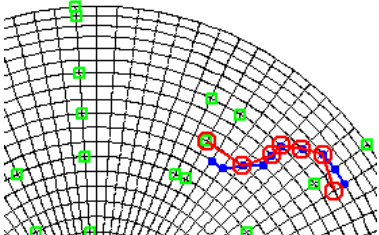
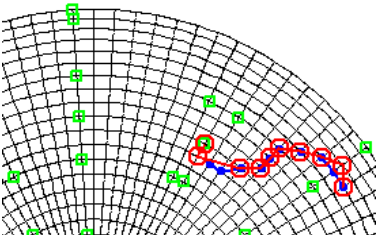

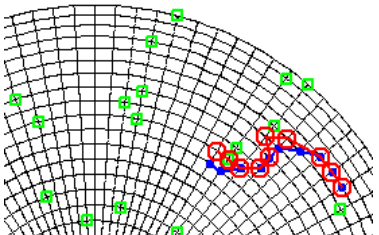
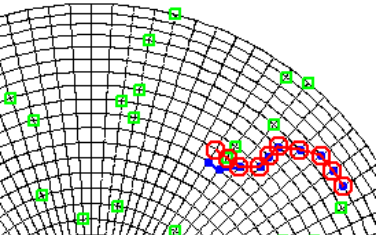
first image of sequence	single pose estimation	sequence estimation
 <p>$\tau = 0.85$</p>	 <p>mean error: 3.11°</p>	 <p>mean error: 1.44°</p>
 <p>$\tau = 0.85$</p>	 <p>mean error: 2.54°</p>	 <p>mean error: 1.59°</p>
 <p>$\tau = 0.85$</p>	 <p>mean error: 1.44°</p>	 <p>mean error: 0.72°</p>

Figure 8.3: Single Pose and Sequence Estimation for Non-Degraded Images. For the single pose estimation the estimated poses (dark gray circles) deviate only little from the true poses (black dots), because the rough estimation already provides the correct view bubble. This means, that almost no outliers occur in the estimated sequence and most improvements achieved by the sequence estimation are due to the fine-tuning mechanism.

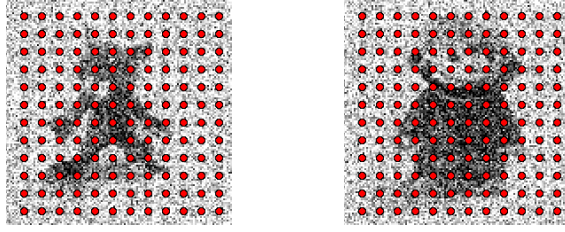


Figure 8.4: Graphs Used For Noisy Test Images. As the segmentation algorithm is not applicable to noisy images it is necessary to cover almost the entire test image by a representing graph.

8.3 Adding Noise

From the good results of the pose estimation from non-degraded images it can be expected that the view-based approach proposed in this thesis even allows pose estimation from images which are degraded by added noise. This problem is discussed in this section.

8.3.1 Methods

The original images of the three test sequences introduced in subsection 8.1.1 are degraded independently by adding Gaussian white noise of zero mean and variance $\mathbf{V} = 0.06$. Examples for noisy images are depicted in figure 8.5.

To gain a representation of the object view displayed in the test image a graph has to be extracted. For non-degraded test images the image has been segmented into object and background before a graph has been extracted only from the object segment. However, for a noisy image a reasonable segmentation into object and background cannot be expected. Thus, to simulate a realistic task in which no information about the object is provided it is necessary to cover almost the entire test image by a graph. In figure 8.4 such graphs are displayed on two of the test views. I chose a rectangular grid graph with equidistant vertices starting with the upper left vertex at pixel position $(10, 10)$ with a distance of 10 pixels in x - as well as y -direction. The graphs used do not cover the entire image, rather, a small area at the boundaries of the image is left blank for the global move of the matching algorithm described in section 3.6. For such a graph \mathcal{G}_T of a test view T the Gabor features are extracted at its vertices. Finally, \mathcal{G}_T is matched on the center images $I_i, i \in R$, of the view bubbles of the object representation \mathcal{R} as described for non-degraded test images in subsection 8.1.1.

To estimate the amount of information on the object's appearance which is available in such a representation of a degraded image I have exemplarily reconstructed two views from their representing graphs as described in subsection 7.1.2. The results are depicted in figure 8.5 as well and reveal a small amount of information on details of the object.

The same single pose and sequence estimation experiments are performed as described in subsections 8.1.1 and 8.2.1 for non-degraded images, this time with the noisy images of the three test sequences. The mean estimation errors are determined for each partitioning of the viewing hemisphere and for both objects.

mean errors, noisy images			
	τ	single pose estimation	sequence estimation
object "Tom"	0.75	64.42°	50.13°
	0.8	40.02°	45.69°
	0.85	34.69°	33.96°
	0.9	29.64°	28.97°
	0.95	11.28°	9.99°
object "dwarf"	0.75	41.79°	36.77°
	0.8	30.98°	23.05°
	0.85	13.48°	12.09°
	0.9	3.9°	4.06°
	0.95	2.09°	1.91°

Table 8.2: Mean Estimation Errors for Degraded Images. Each value given in degree is the mean estimation error computed from 30 noisy test views. The first column lists the mean errors which occur if each view has been estimated independently. In the second column the mean errors are presented for the sequence estimation experiments.

8.3.2 Results

The mean estimation errors of the single pose and sequence estimation for degraded images are summarized in table 8.2. As expected, the mean errors for noisy images are larger than for non-degraded test views for both objects and each partitioning of the viewing hemisphere. Again, the estimation results improve with increasing values of τ , i.e., with a larger number of sample views in the object representation, for both objects and for single as well as sequence estimation. However, there is a difference to the results obtained from non-degraded images. The mean estimation errors for the "dwarf" object are smaller than those obtained for the "Tom" object.

Concerning the question of interest, for the majority of the values (with one exception for each object) the mean estimation errors are smaller for the sequence estimation than for the single pose estimation. However, this difference is not significant in any case.

In contrast to the experiments carried out with non-degraded images, here the improvements achieved by the aid of the additional neighborhood information are also due to the outlier mechanism of the sequence estimation algorithm, instead of the fine-tuning mechanism, because the rough estimation of a view of a test sequence more often fails for degraded images. This is particularly obvious for object representations which contain only a few sample views. This effect is demonstrated in the first two examples in figure 8.6. A complete sequence with outlier correction is displayed in figure 8.7. Minor improvements due to the fine-tuning mechanism occur mainly if object representations are

used which contain a larger number of sample view. Examples for that are shown in the last two rows of figure 8.6 and in the figures 8.8 and 8.9.

8.3.3 Discussion

The quality of the pose estimation from test views degraded by noise for a reasonable partitioning of the viewing hemisphere can be regarded as fairly good. This holds especially for the “dwarf” object where the mean estimation errors are less than 15° . For individual test sequences the proposed estimation methods are capable to provide mean estimation errors less than 4° even if the used object representation contains only few sample views as displayed in the figures 8.7 and 8.8.

The better estimation results obtained for the “dwarf” object than for the “Tom” object can be explained by the fact that the addition of noise has a larger impact on the “Tom” object. For example, the face of “Tom” is characterized by a larger amount of fine details, i.e., high frequencies, than the face of the “dwarf”. As high frequencies are more affected by the addition of noise than low frequencies the available amount of information about the “Tom” object is smaller than that of the “dwarf” object after the addition of noise. The larger amount of low frequencies contained in the views of the “dwarf” can still be utilized for view discrimination. This effect is demonstrated in figure 8.5, where the observer can recognize the “dwarf” reconstructed from the noisy image easier than the reconstructed “Tom”.

For noisy test images the rough estimate of the single pose estimation algorithm more often provides a wrong view bubble than for non-degraded test images. In this case the fine-tuning mechanism of the sequence estimation algorithm is usually ineffective. In addition, the outlier mechanism can only improve the estimation of the sequence if *single* views have been misestimated with a large distance to the true pose. For degraded images and a very sparse object representation, however, the rough estimation for more than one successive test view fails. This might be a reason for the *non-significant* improvement provided by the sequence estimation. The benefit of utilizing neighborhood information of successive views could however be enhanced by a more elaborate algorithm which does not only take the best estimates inside *one* view bubble into account, but for example, tests also views in neighboring view bubbles.

Concluding I can say, that the proposed view-based approach, which includes the sparse object representation \mathcal{R} as well as the techniques for generating representations $\hat{\mathcal{G}}$ of unfamiliar views, seems to be suitable to perform perception tasks, demonstrated here for the estimation of an object’s pose, even if the amount of information on the object provided in the test images is reduced considerably by added noise. For test images degraded by noise the utilization of the temporal context of successive views, which is presented to an observer by a rotating object, can slightly improve the pose estimation.


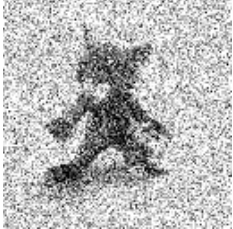

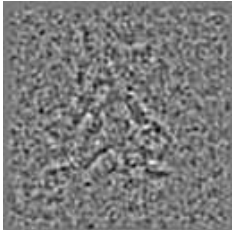




	non-degraded images	noisy images
test images for pose and sequence estimation		
reconstructed from Gabor filter responses		
test images for pose and sequence estimation		
reconstructed from Gabor filter responses		

Figure 8.5: Reconstructions of Non-Degraded and Noisy Images from Gabor Responses. The original images of the test sequences are degraded by adding Gaussian white noise of zero mean and variance $V = 0.06$ as depicted in the second column. Single pose and sequence estimation experiments are performed for the test sequences degraded by this amount of noise. The reconstructions from their Gabor filter responses illustrate the small amount of information which can be extracted from the degraded images by Gabor wavelets.

noisy images

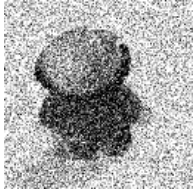
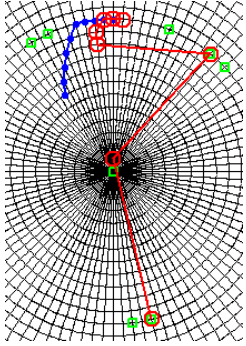
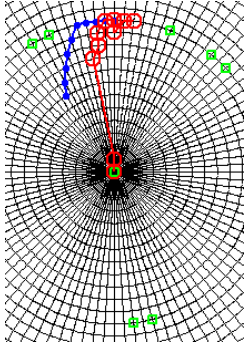
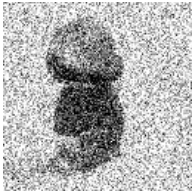
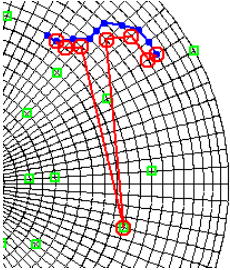
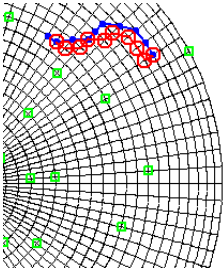
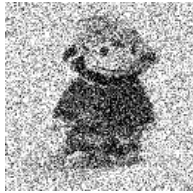
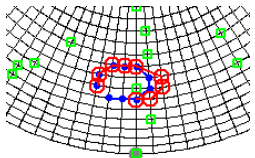
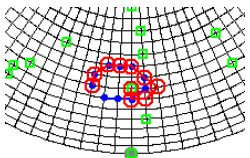

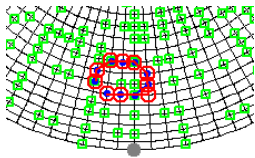
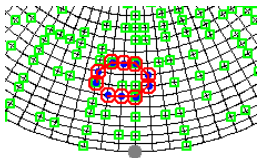
first image of sequence	single pose estimation	sequence estimation
 $\tau = 0.75$	 mean error: 32.46°	 mean error: 17.19°
 $\tau = 0.8$	 mean error: 9.39°	 mean error: 3.18°
 $\tau = 0.85$	 mean error: 2.16°	 mean error: 1.89°
 $\tau = 0.95$	 mean error: 0.36°	 mean error: 0°

Figure 8.6: Single Pose and Sequence Estimation for Noisy Images. For images degraded by added noise the improvements of the sequence estimation in comparison to the single pose estimation are also due to the fact that single outliers of the estimated sequence can be corrected. This is particularly obvious in the first two examples, which have been obtained from object representations with only a few sample views ($\tau = 0.75$ and 0.8).

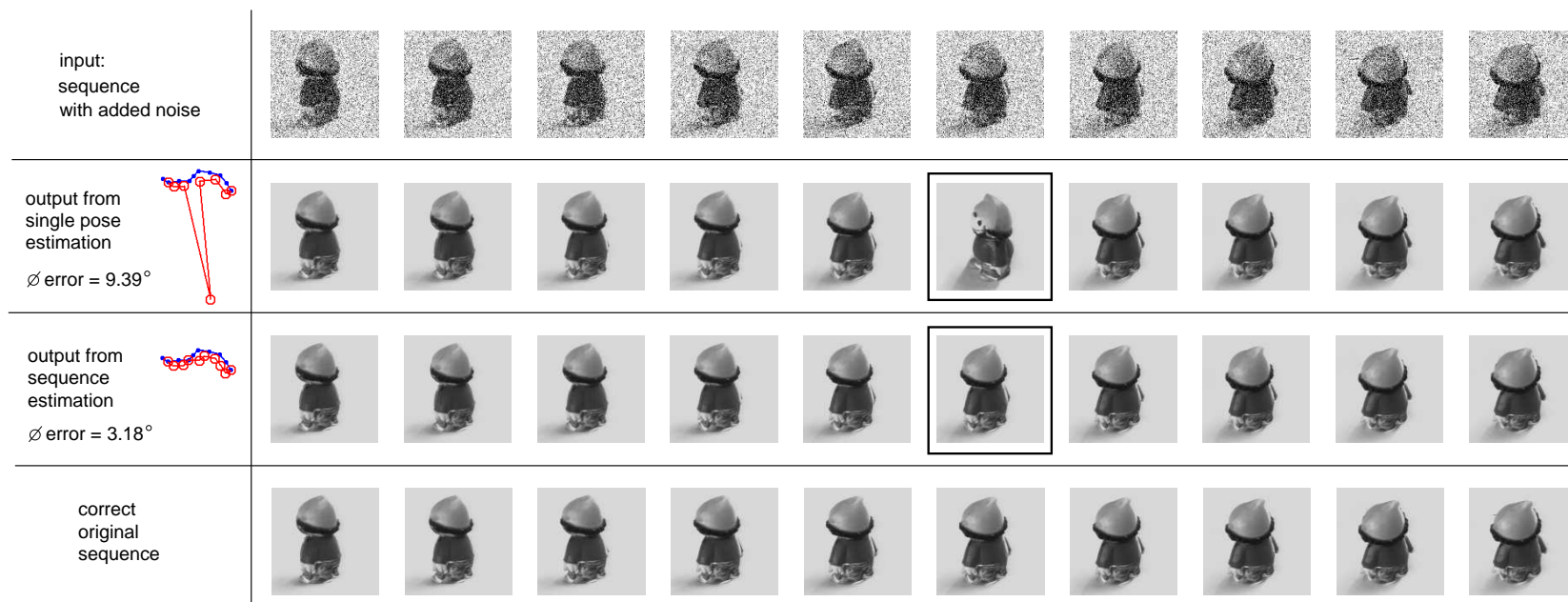
single pose and sequence estimation, object "dwarf", noisy test images, $\tau = 0.8$ 

Figure 8.7: Estimation Example - Object “Dwarf”, $\tau = 0.8$. This is an example of pose estimation with an object representation consisting of only a few sample views ($\tau = 0.8$). For images degraded by noise the failures of the single pose estimation are often due to a misleading result of the first step of the algorithm. The noise often leads to a wrong rough estimate of the view’s pose. Thus, improvements provided by the sequence estimation can be achieved by a correction of single outliers. The outliers, which are corrected by the sequence estimation, are marked in this figure (compare with the second row of figure 8.6).

single pose and sequence estimation, object "dwarf", noisy test images, $\tau = 0.85$

input: sequence with added noise	
output from single pose estimation \emptyset error = 2.16°	
output from sequence estimation \emptyset error = 1.89°	
correct original sequence	

Figure 8.8: Estimation Example - Object “Dwarf”, $\tau = 0.85$. This is an example of pose estimation from noisy images with a reasonable partitioning of the viewing hemisphere. The results from the single pose estimation are already satisfying. Minor improvements can be achieved by the sequence estimation due to the fine-tuning mechanism (compare with the third row of figure 8.6).

single pose and sequence estimation, object "Tom", noisy test images, $\tau = 0.95$

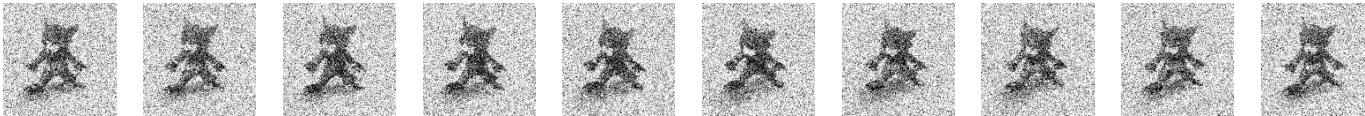

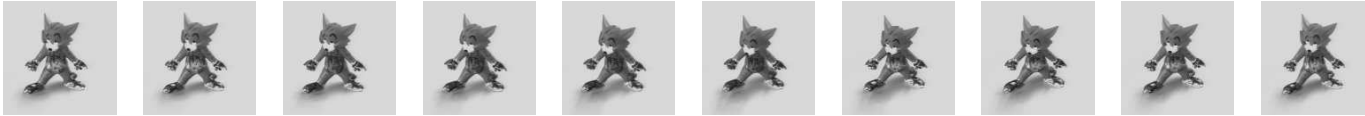

input: sequence with added noise	
output from single pose estimation \emptyset error = 0.36°	
output from sequence estimation \emptyset error = 0°	
correct original sequence	

Figure 8.9: Estimation Example - Object "Tom", $\tau = 0.95$. This final example shows a perfect result obtained from the sequence estimation algorithm with a large number of views in the object representation. The pose of the third view of the sequence obtained from single pose estimation has been corrected by the fine-tuning mechanism (compare with the last row of figure 8.6).

Chapter 9

Summary and Conclusions

In brain research as well as in computer science three-dimensional object perception is still an unsolved problem. This thesis aimed at studying the possibility of establishing a representation of a three-dimensional object from two-dimensional views only and analyzing the capabilities of such a view-based representation to perform perception tasks such as estimating the pose of an object. For this purpose I simulated a biologically plausible system which is able to independently learn sparse, view-based object representations from sample views of real-world objects. Representations of unfamiliar views can be derived from those of stored views by a linear combination of object point positions and an interpolation of object point features. The system is capable of estimating object poses, given a sufficient number of sample views, even from input images degraded by added noise, which proves the capability of the view-based approach to perform perception tasks.

In detail, my investigations were guided by the questions **Q1** to **Q4** specified in the introduction. Answers to these questions provided by the current state of research in disciplines such as biology, psychology, engineering, and computer science are already listed in the summary of chapter 2. Here the results and conclusions I obtained from my simulations are briefly summarized.

Q1 First of all, the general question about a view-based representation of three-dimensional objects can be affirmed. It is possible to interpret the visual perception of objects by a purely view-based approach. The establishment of an explicit, three-dimensional model of an object does not seem to be necessary to perform perception functions. That means that the main thesis of this work put forward in the introduction (chapter 1) can be supported.

Q2 For each view of an object I determined a surrounding area of viewpoints for which the appearance of the object changes only slightly. The determination of such areas for each view on the viewing hemisphere of an object reveals views which possess larger areas of pose robustness than the majority of other views. These views can be regarded as *canonical views*. The importance of canonical views for human object perception has been suggested by many physiological and psychological studies. Furthermore, I chose only some of these areas of pose robustness which are sufficient to cover the whole viewing hemisphere. The representations of these selected areas account for the view-based object representation. Each selected area can be regarded as an *aspect*, which is another well-known concept supported by behavioral studies.

- Q3** The distribution of the views which constitute the representation of a three-dimensional object in my simulations depend on the object and on the precision with which unfamiliar views must be recoverable from stored views. The range of generalization obtained for a reasonable partitioning of the viewing hemisphere is in accordance with psychological and physiological results. On average this range, within which an inference from a familiar to an unfamiliar view is possible, comprises views with a distance of about 30° to 40° .
- Q4** In my approach representations of unfamiliar views are derived from those of familiar views by a *linear combination* of object point positions and a biologically inspired *interpolation* of object point features. The necessary *correspondences* between the stored views are provided by tracking object points from view to view. This procedure has its parallels in psychology as well. From the quality of the resulting representations of unfamiliar views reasonable recognition rates can be expected, the more so as the possibility has been demonstrated to generate morphed versions of unfamiliar views of fair quality with the linear combination approach.

The above-mentioned viewpoint-invariant *recognition* of objects is a point which is directly connected with my work but has not been analyzed in this thesis. It requires a large data base with images for several objects. Roughly sketched, it could function similarly to the pose estimation, with the difference that the representing graph of the test view is matched with the representations of *all* objects stored in the data base instead of *one* object only. As an alternative, object recognition could be realized by generating morphed versions of unfamiliar views of objects from views stored in their representations and then matching the graph of the test view on the morphed views. This latter method could also be tried for pure pose estimation.

The necessity to provide a large number of object views together with their relative positions in order to *learn* a sparse representation of the object does not seem to have a counterpart in the functionality of our brains and represents a limitation of this work in the current stage. Thus, a next step of investigation could be the learning of a view-based object representation following insights from the biological sciences. For instance, the active manipulation of objects can be an important source of information during the acquisition of an object representation, because the information about the relative positions of views would be provided by the joint angles of the arm which rotates the object. To avoid the complete supply of object views it can be advantageous to integrate a motor-controlled feedback loop which actively decides which aspects of an object require a further inspection.

Besides the biological relevance of this thesis, which is its major concern, there are nevertheless technical applications as well. For example, if the generation of unfamiliar views from some sample views is feasible (as shown in chapter 6), then the rendering of whole objects is possible from only some stored reference views, which could be relevant for data compression.

Appendix A

Sequences of Matched and Tracked Local Object Features

The figures A.1 to A.6 display parts of sequences of matched and tracked object points and the referring similarity diagrams. In the figures A.7 to A.14 the complete sequences are displayed. See subsection 4.3.2 for a description.

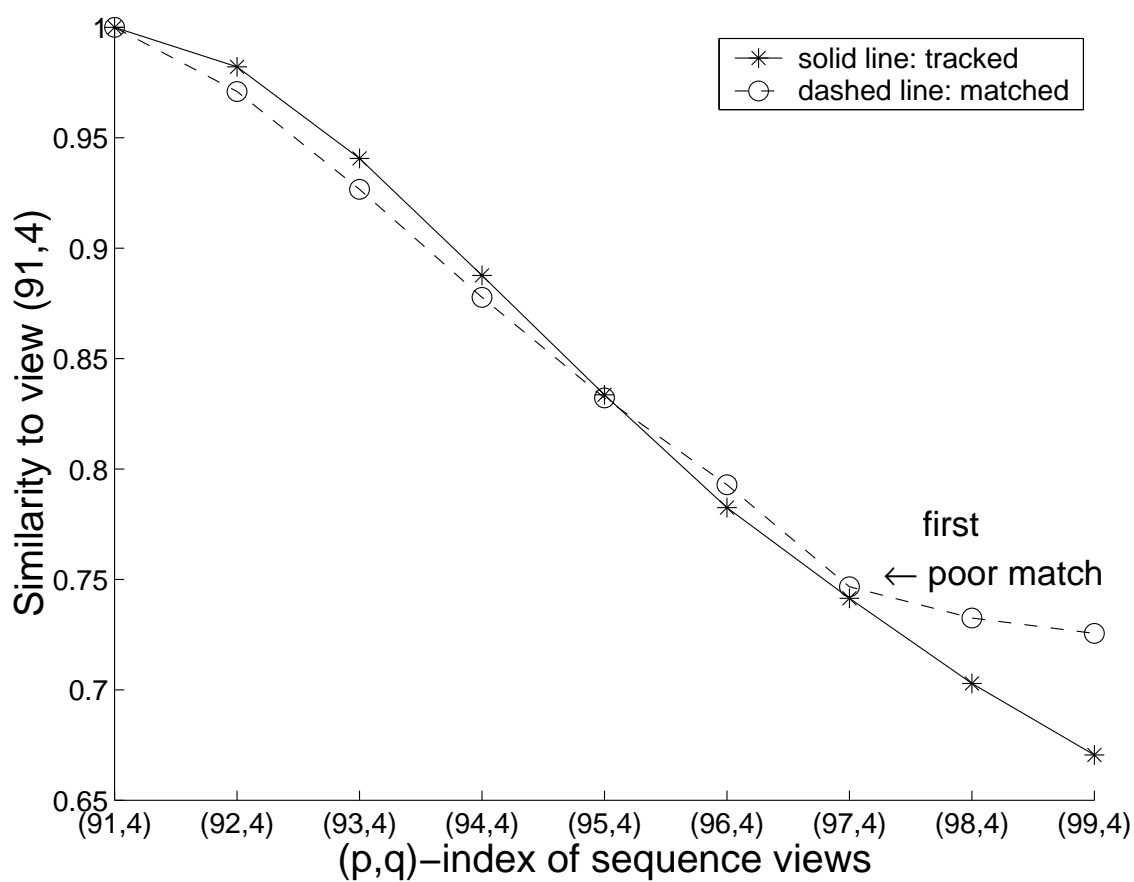
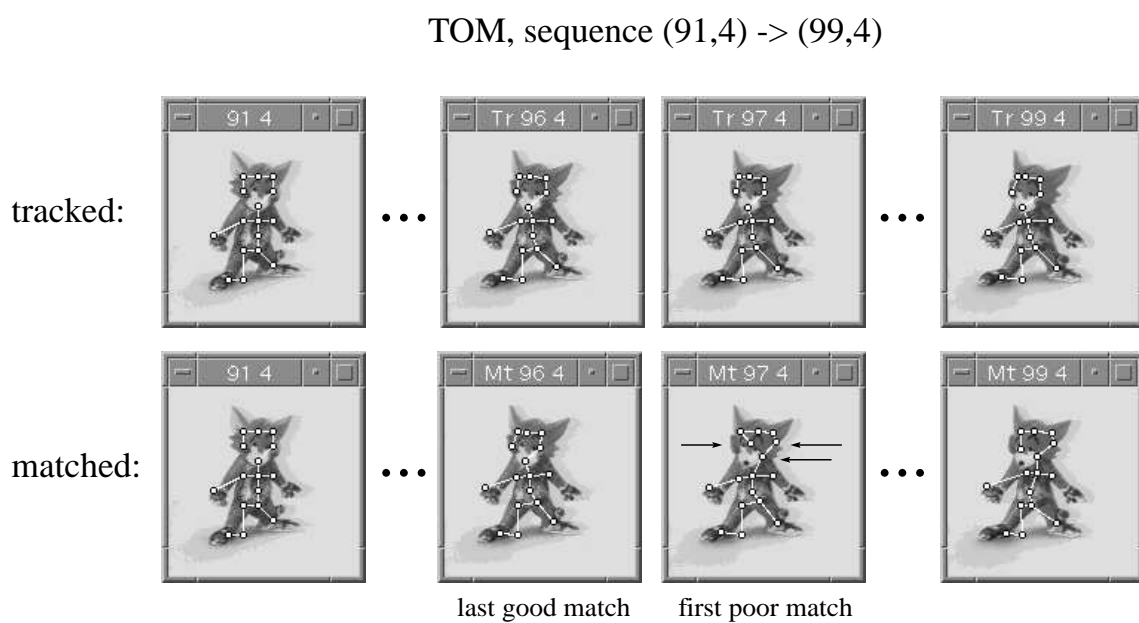


Figure A.1: Object "Tom", Second Sequence With Similarity Diagram.

TOM, sequence (40,6) \rightarrow (49,6)

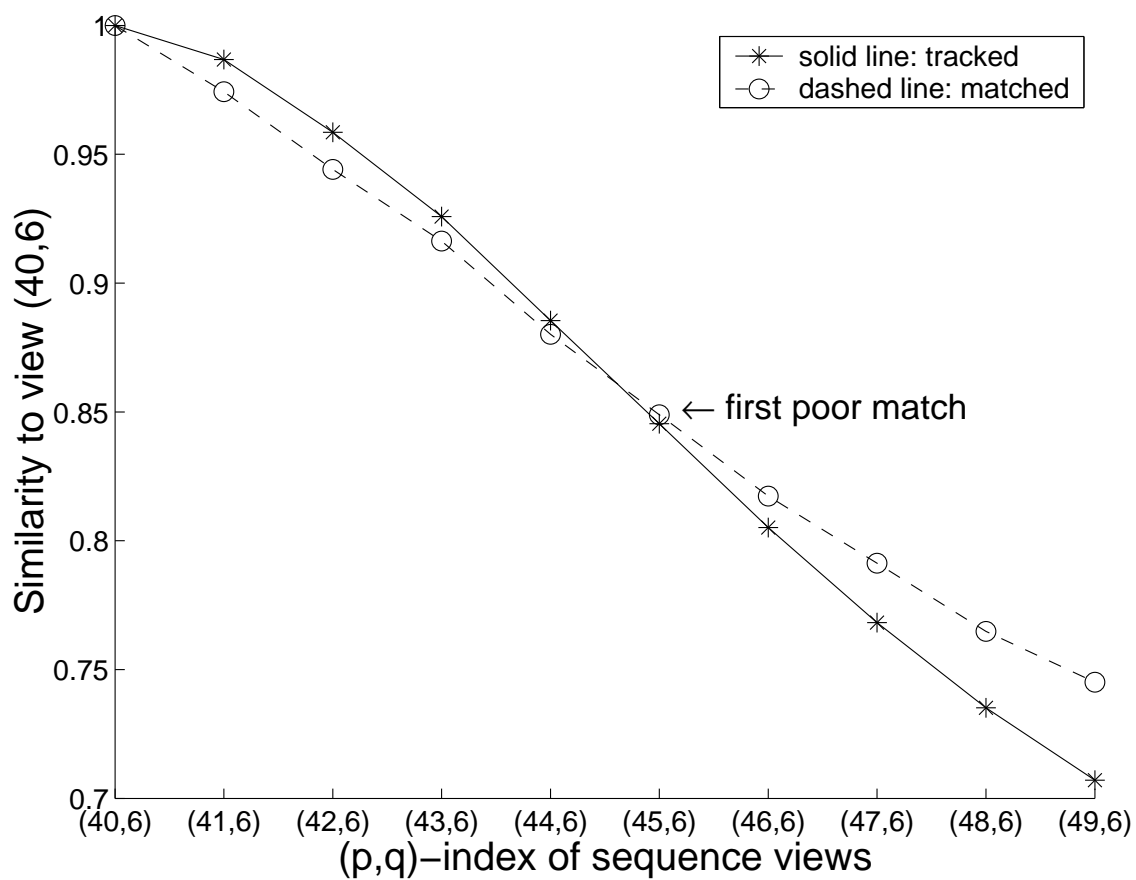
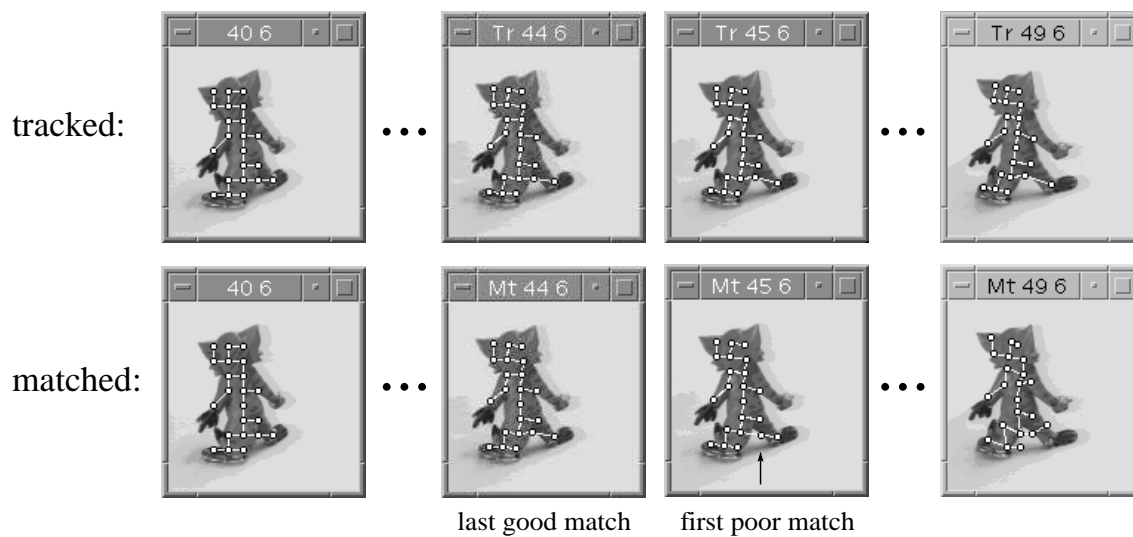


Figure A.2: Object "Tom", Third Sequence With Similarity Diagram.

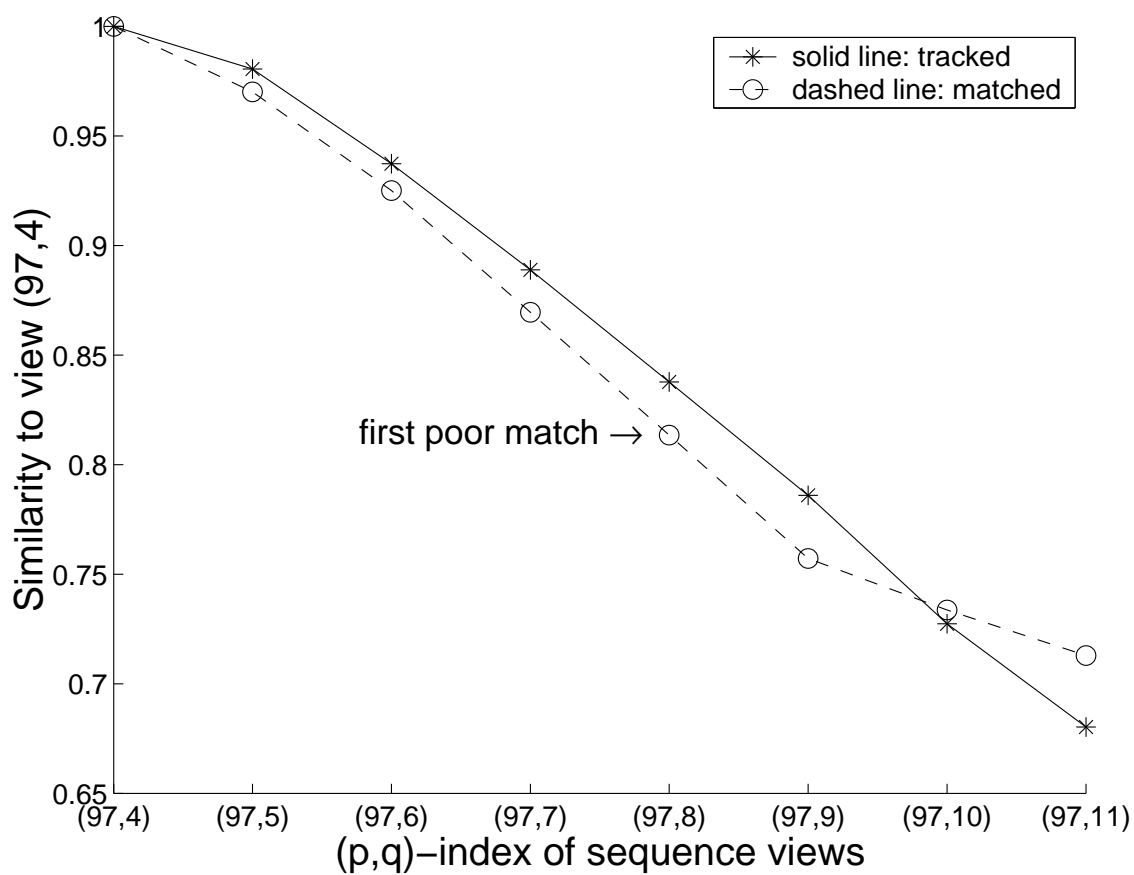
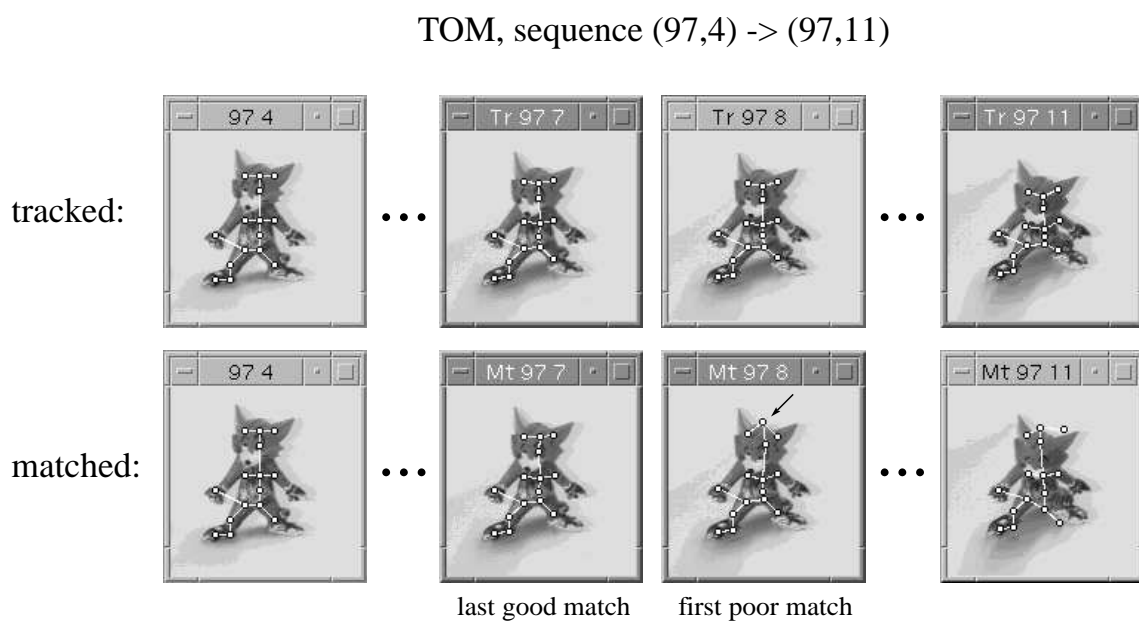


Figure A.3: Object "Tom", Fourth Sequence With Similarity Diagram.

DWARF, sequence (80,6) -> (80,12)

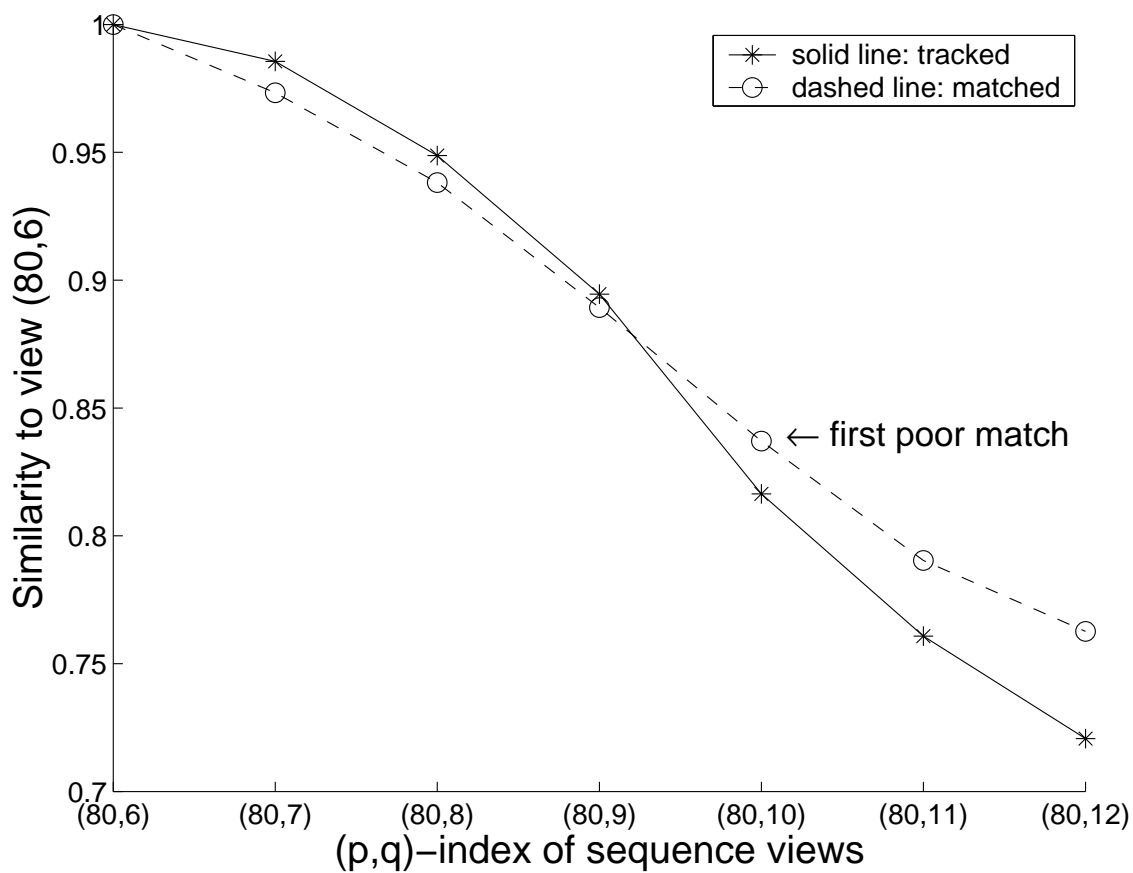
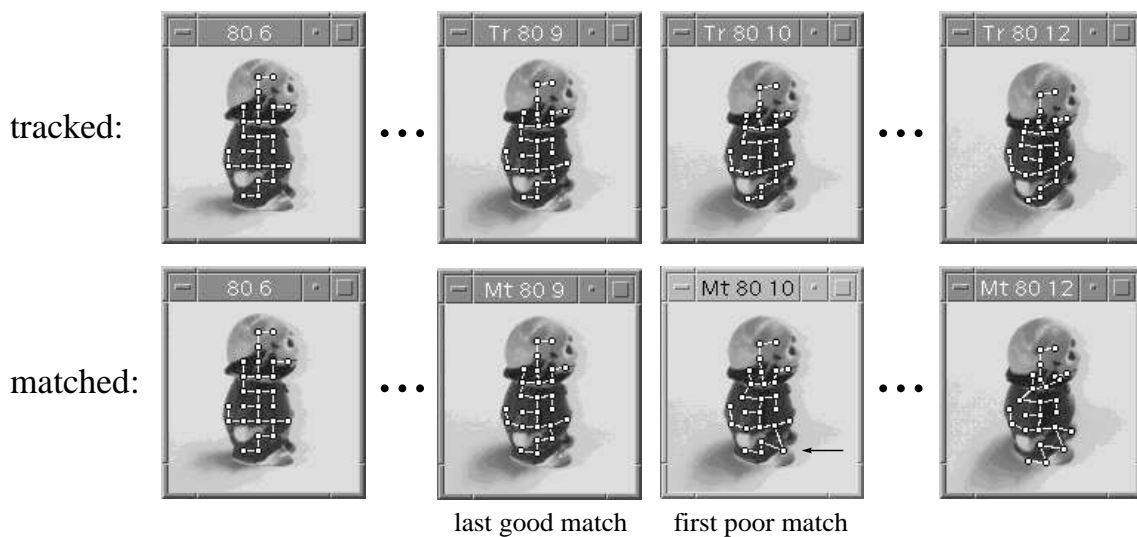


Figure A.4: Object “Dwarf”, Second Sequence With Similarity Diagram.

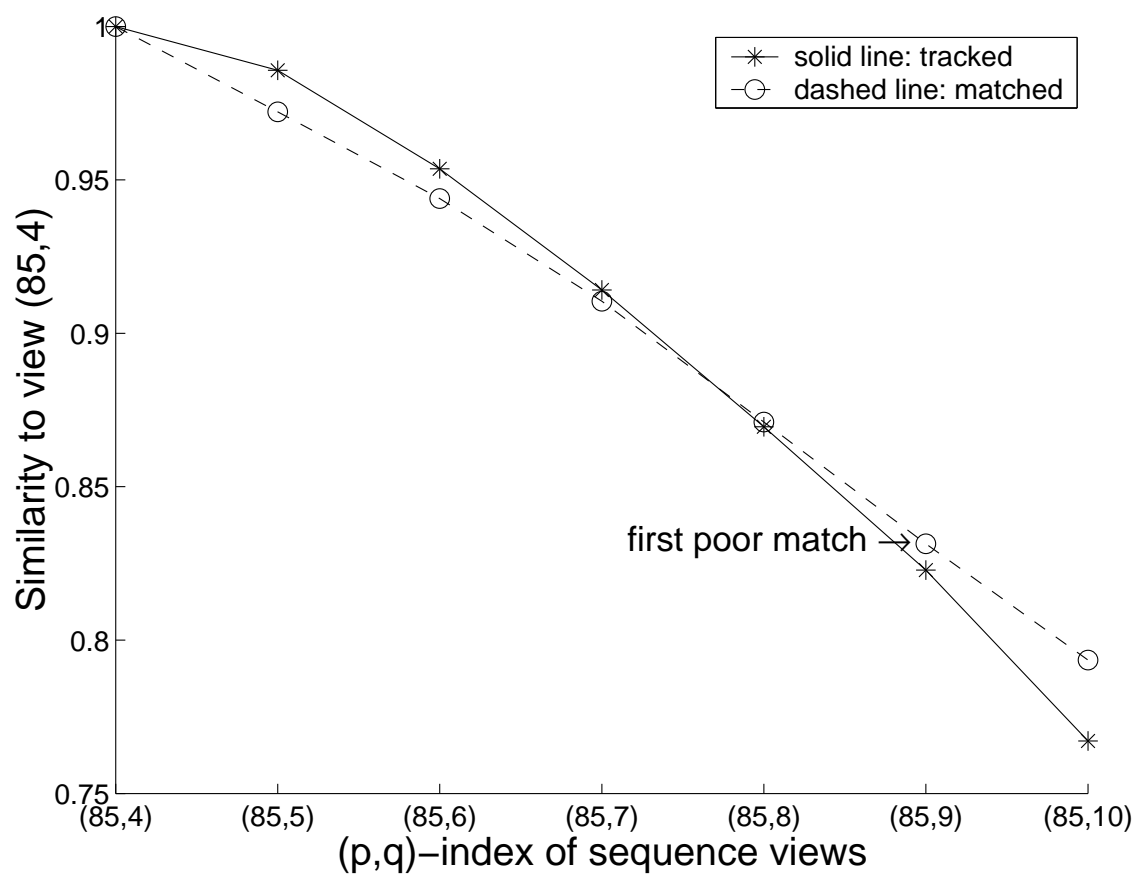
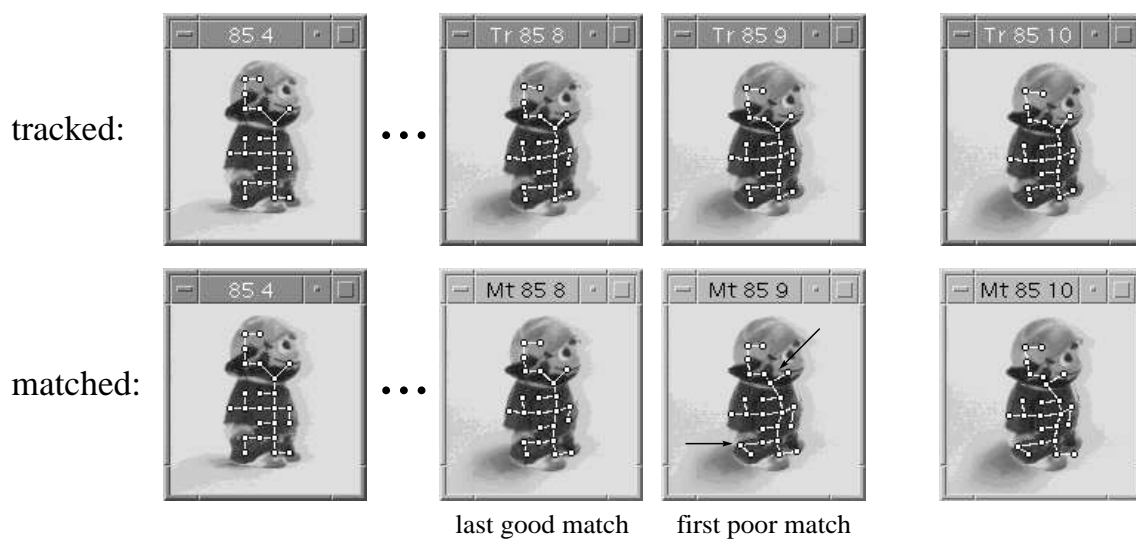
DWARF, sequence (85,4) \rightarrow (85,10)

Figure A.5: Object "Dwarf", Third Sequence With Similarity Diagram.

DWARF, sequence (64,18) -> (71,18)

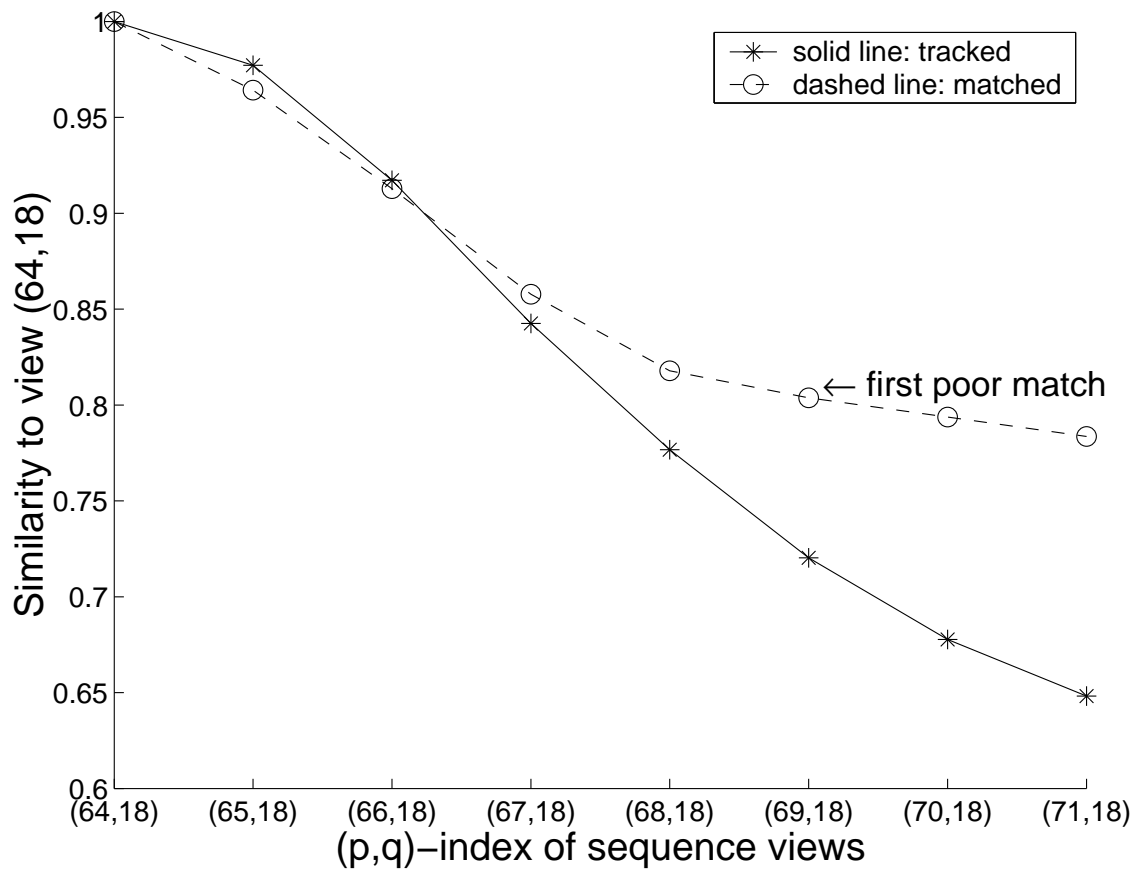
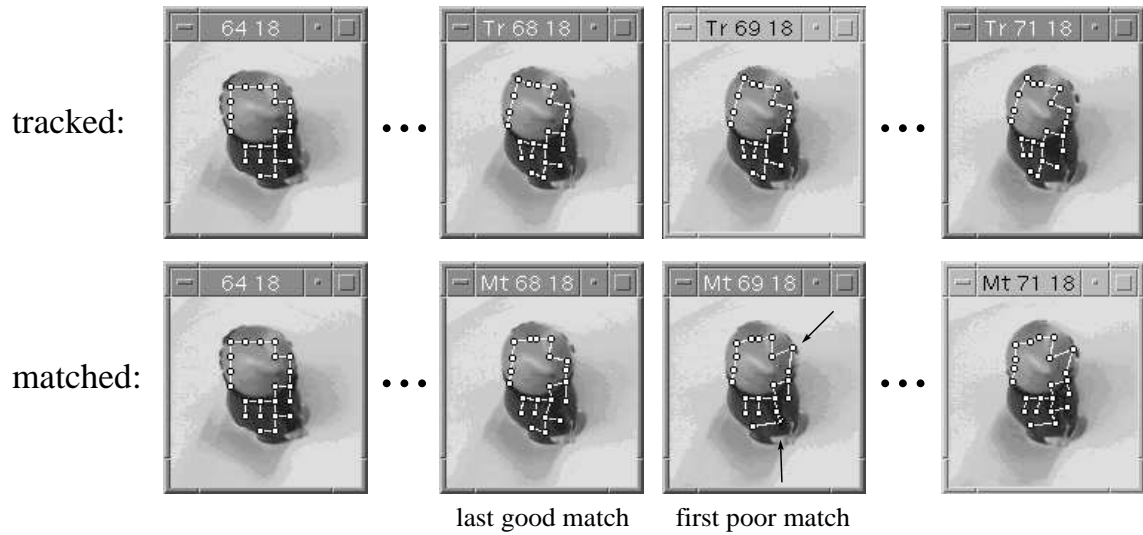
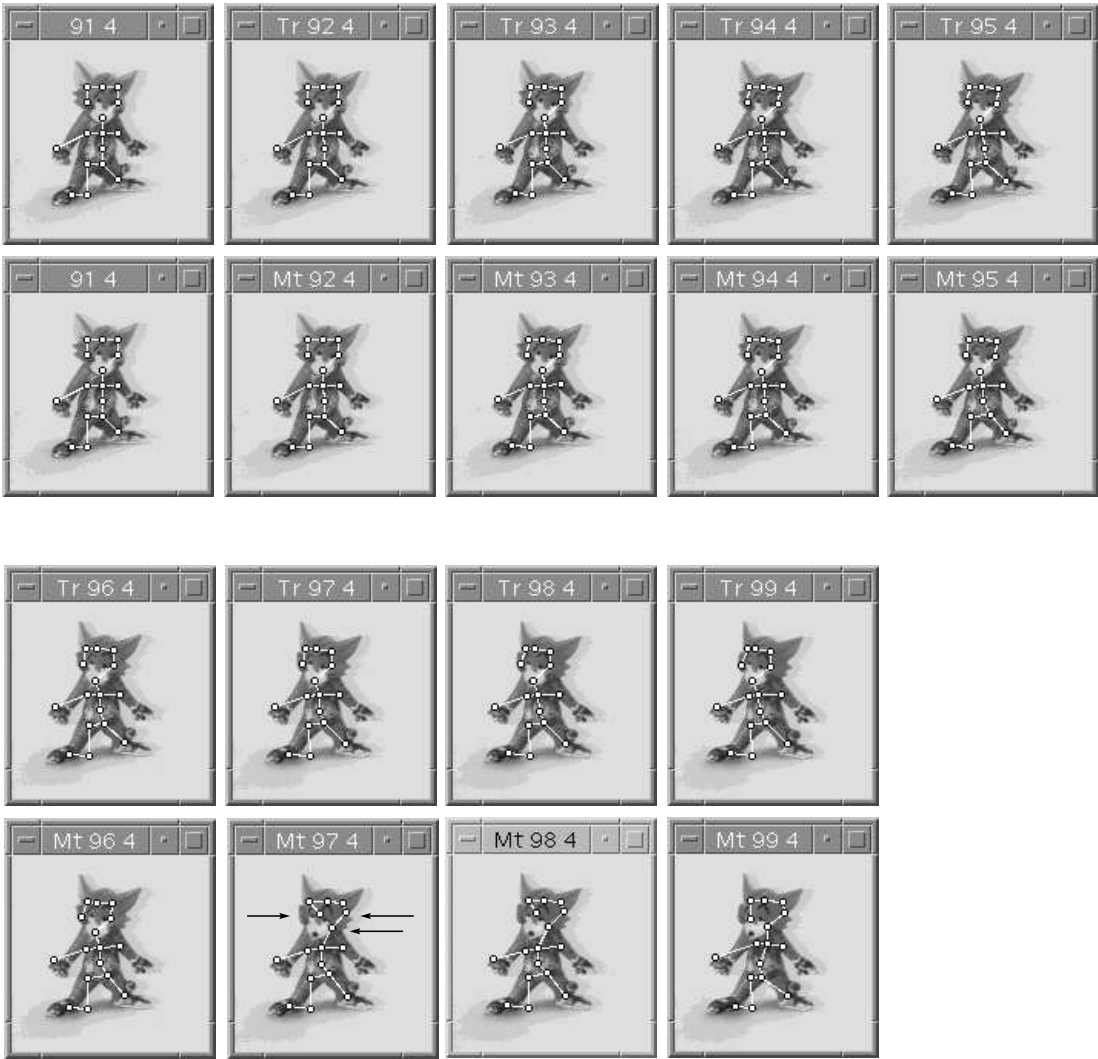


Figure A.6: Object "Dwarf", Fourth Sequence With Similarity Diagram.

TOM, sequence (48,5) \rightarrow (36,5)

Figure A.7: Object "Tom", Complete First Sequence.

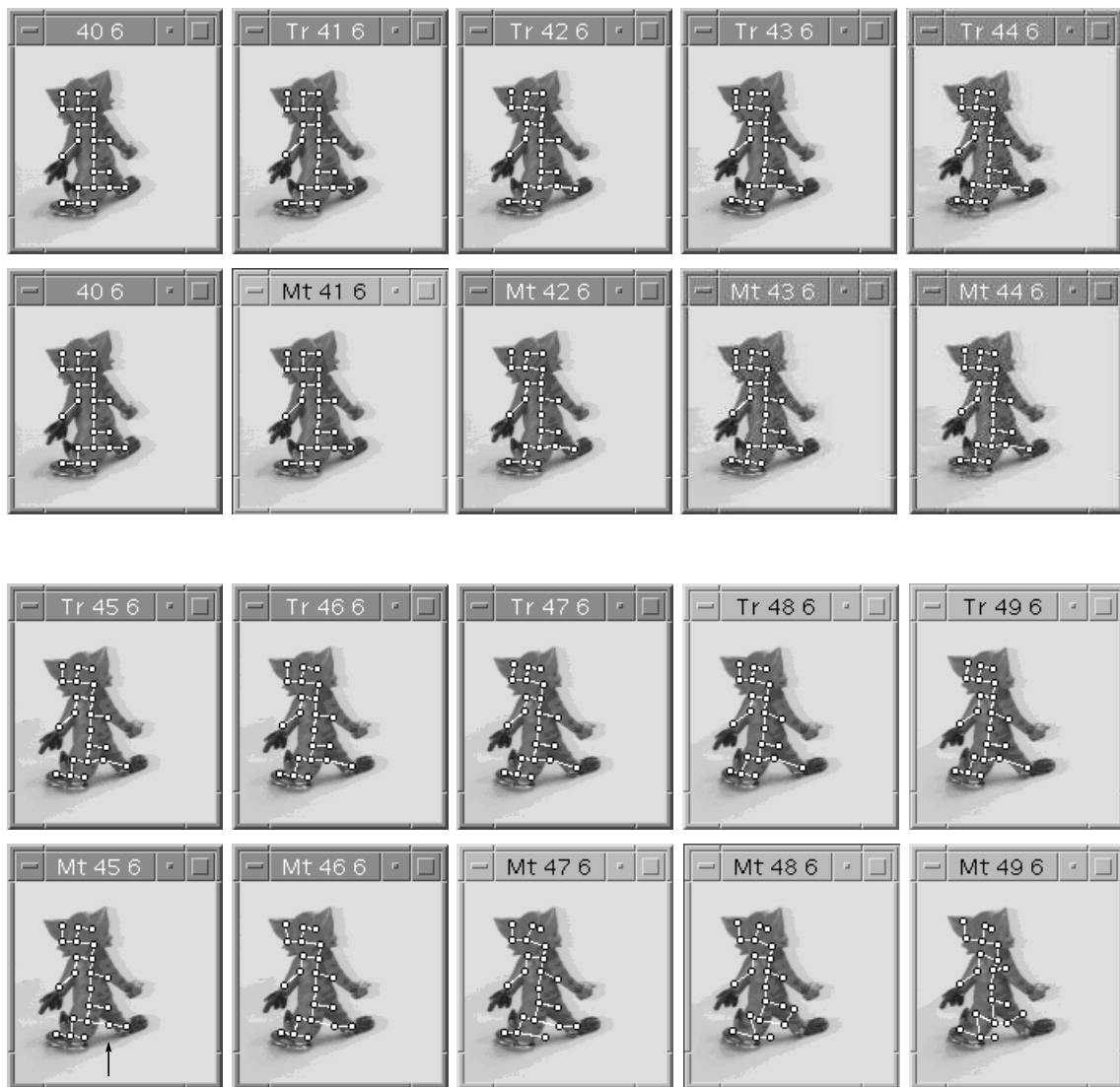
TOM, sequence (91,4) -> (99,4)



first poor match

Figure A.8: Object “Tom”, Complete Second Sequence.

TOM, sequence (40,6) -> (49,6)



first poor match

Figure A.9: Object “Tom”, Complete Third Sequence.

TOM, sequence (97,4) -> (97,11)

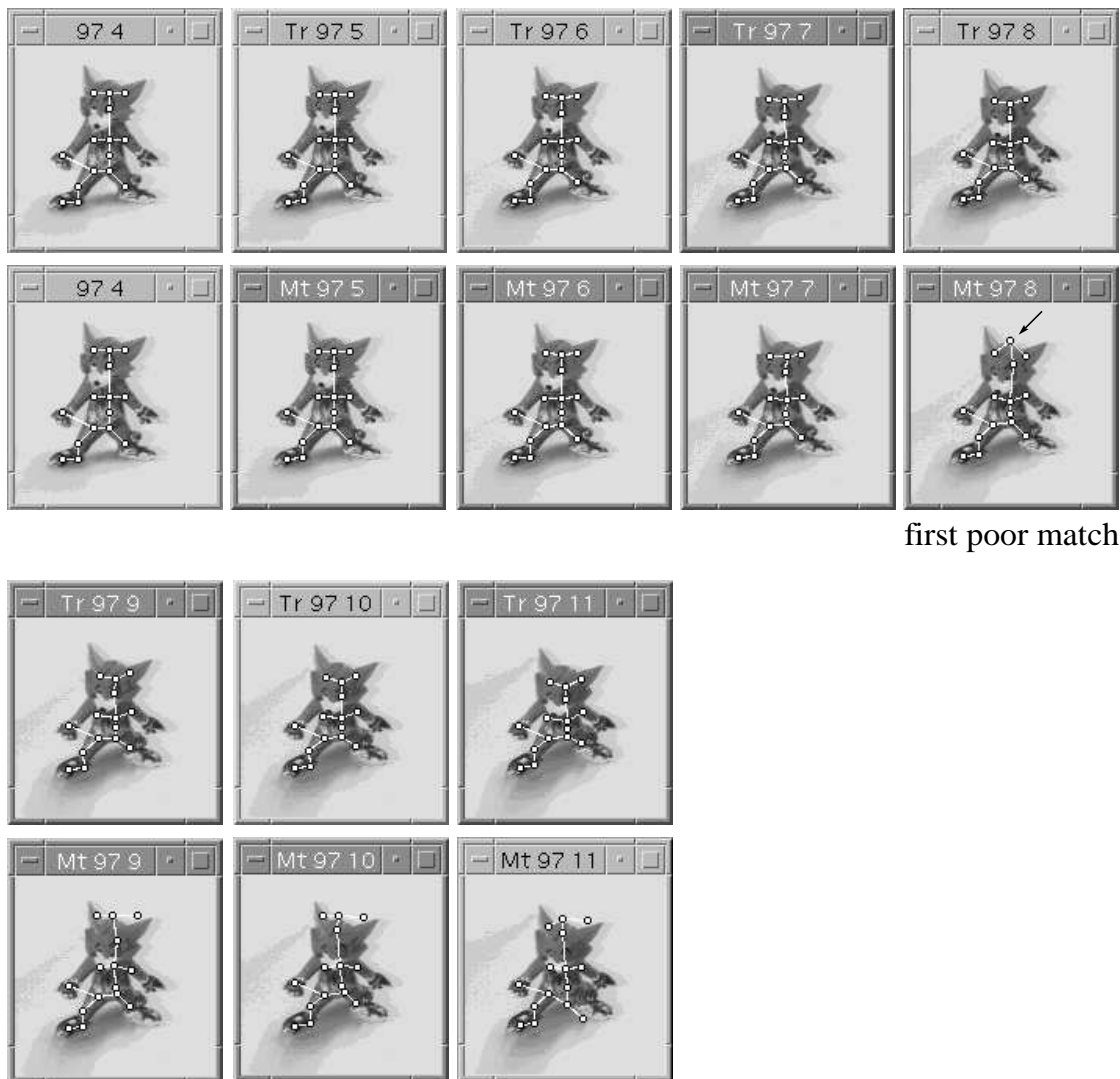
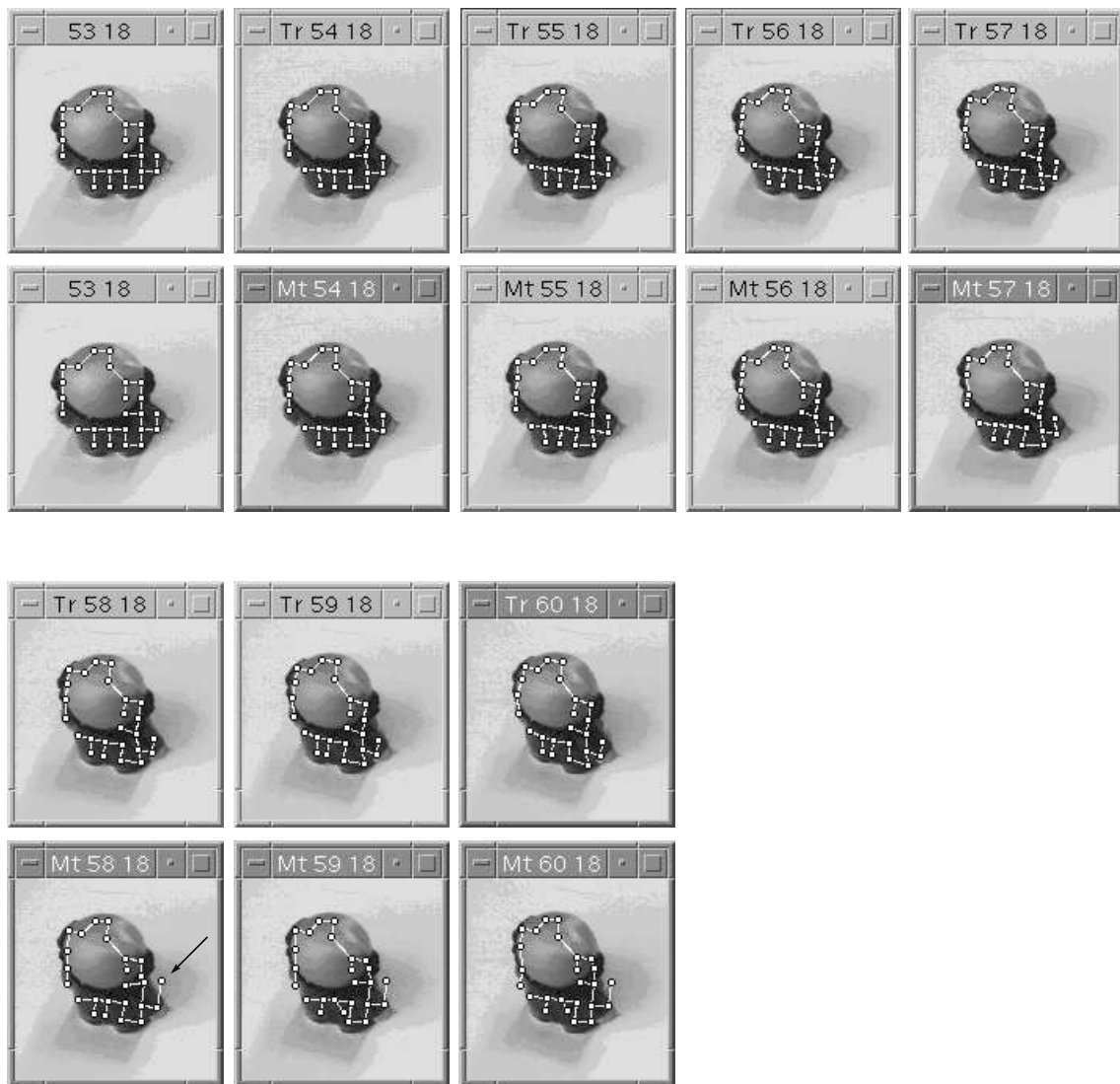


Figure A.10: Object "Tom", Complete Fourth Sequence.

DWARF, sequence (53,18) \rightarrow (60,18)

first poor match

Figure A.11: Object "Dwarf", Complete First Sequence.

DWARF, sequence (80,6) -> (80,12)

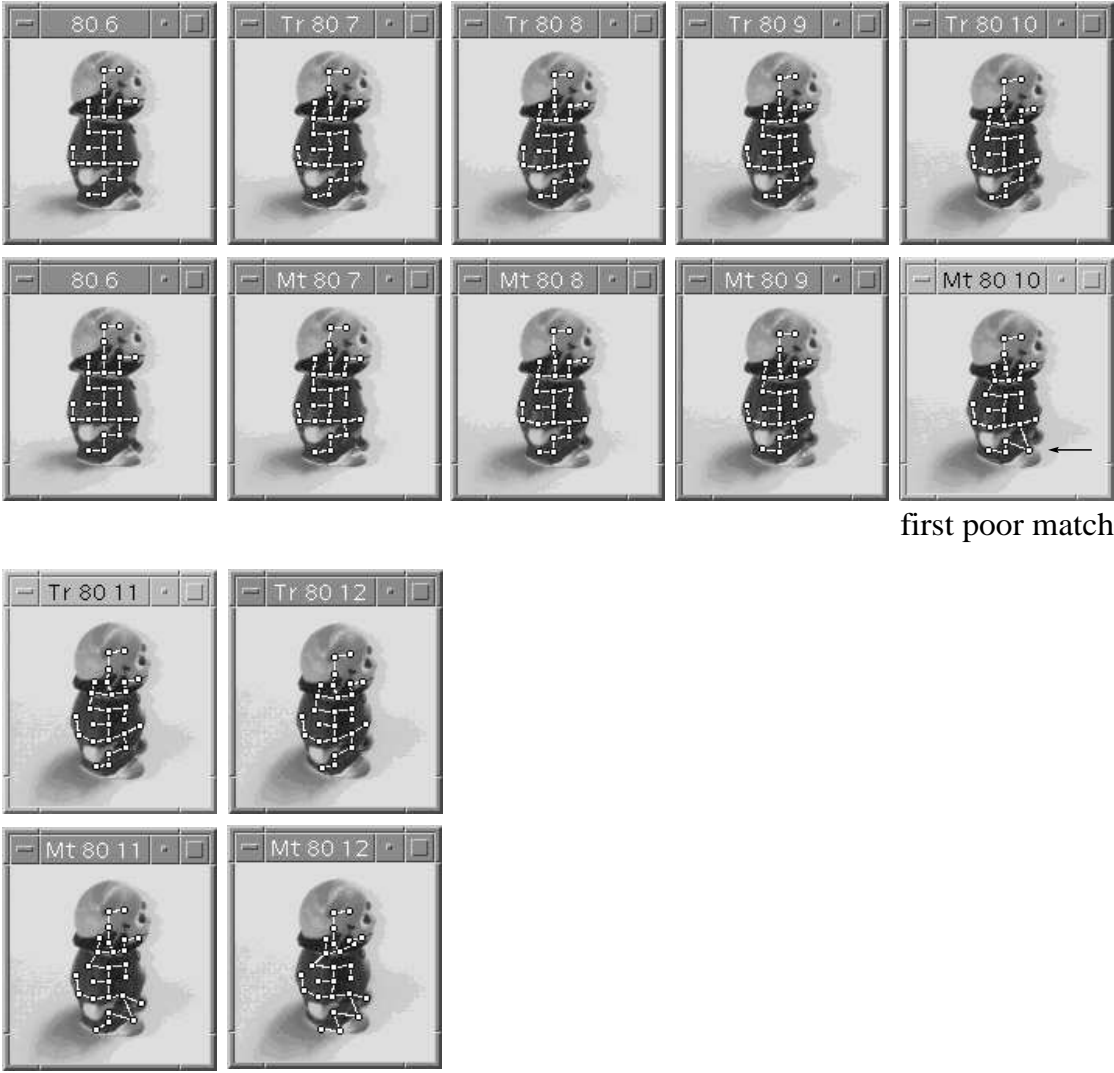
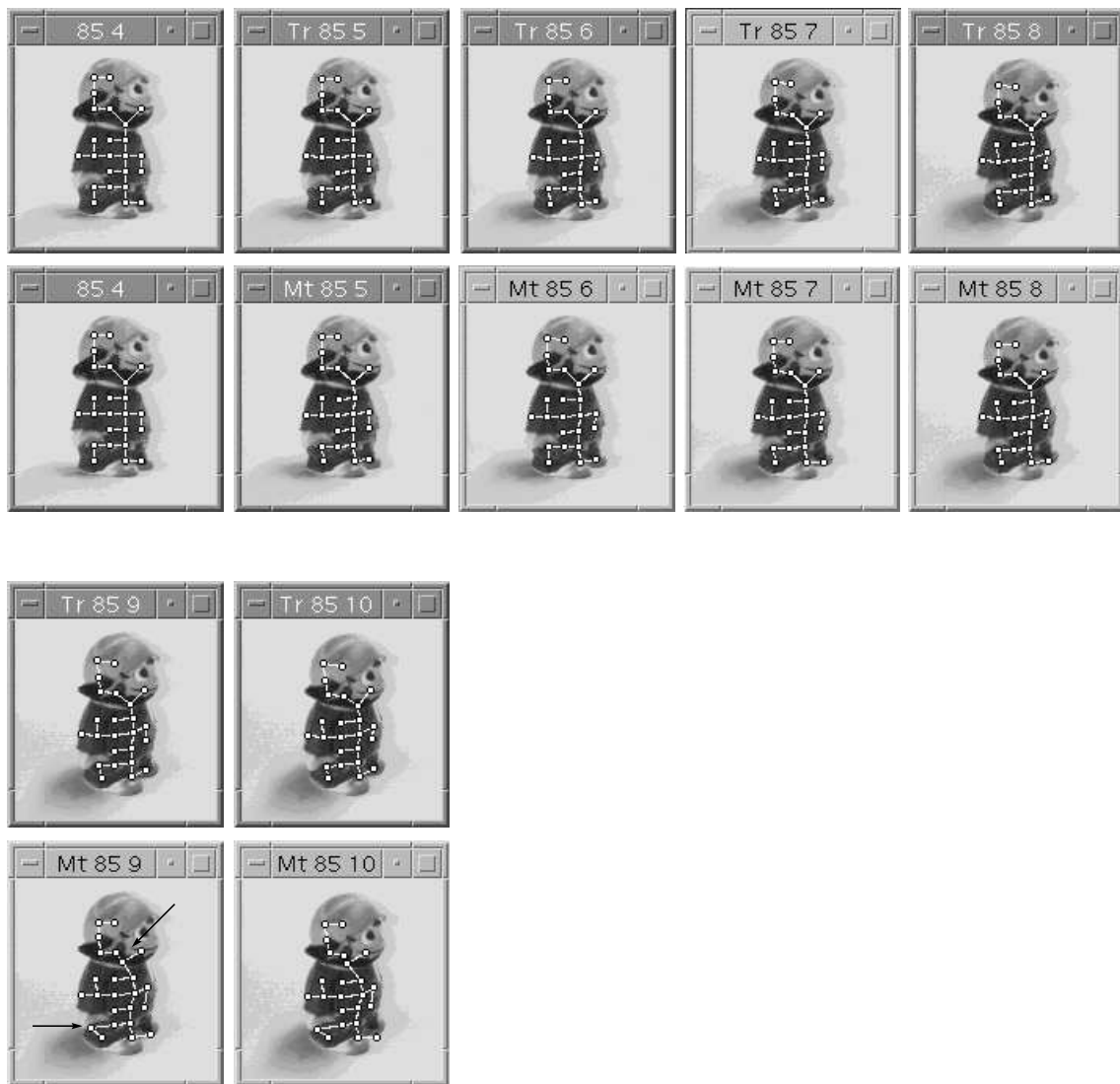


Figure A.12: Object “Dwarf”, Complete Second Sequence.

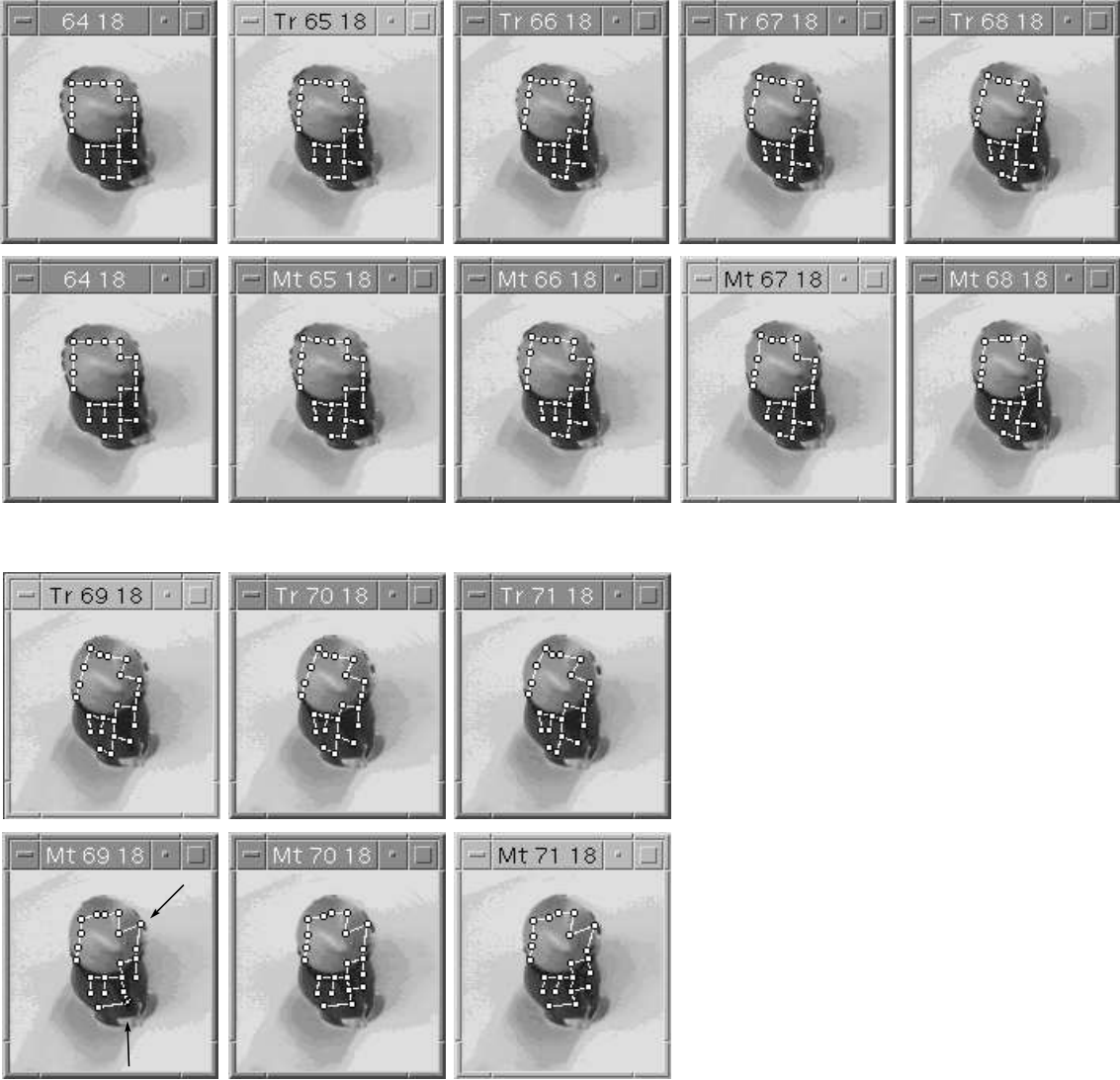
DWARF, sequence (85,4) -> (85,10)



first poor match

Figure A.13: Object "Dwarf", Complete Third Sequence.

DWARF, sequence (64,18) -> (71,18)



first poor match

Figure A.14: Object “Dwarf”, Complete Fourth Sequence.

Appendix B

Linear Combination of Object Point Positions

The position of a view on the viewing hemisphere can be defined by its pan and tilt angles φ and λ with

- (i) $0 \leq \varphi < 2\pi$ and
- (ii) $0 \leq \lambda \leq \frac{\pi}{2}$.

(see figure 6.3). If the original coordinate system spanned by the axes \mathbf{X} , \mathbf{Y} , and \mathbf{Z} is rotated first with the angle φ around the \mathbf{Y} -axis, secondly with the angle $-\lambda$ around the newly risen \mathbf{X} -axis, then the resulting total rotation can be described by the following matrix:

$$ROT(\varphi, \lambda) = Rot(\mathbf{Y}, \varphi) \cdot Rot(\mathbf{X}, -\lambda) \quad (\text{B.1})$$

with the rotation matrices

$$Rot(\mathbf{Y}, \varphi) = \begin{pmatrix} \cos(\varphi) & 0 & \sin(\varphi) \\ 0 & 1 & 0 \\ -\sin(\varphi) & 0 & \cos(\varphi) \end{pmatrix}, \quad (\text{B.2})$$

$$Rot(\mathbf{X}, -\lambda) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\lambda) & \sin(\lambda) \\ 0 & -\sin(\lambda) & \cos(\lambda) \end{pmatrix}, \quad (\text{B.3})$$

and

$$ROT(\varphi, \lambda) = \begin{pmatrix} \cos(\varphi) & -\sin(\varphi)\sin(\lambda) & \sin(\varphi)\cos(\lambda) \\ 0 & \cos(\lambda) & \sin(\lambda) \\ -\sin(\varphi) & -\cos(\varphi)\sin(\lambda) & \cos(\varphi)\cos(\lambda) \end{pmatrix}. \quad (\text{B.4})$$

Let \vec{o} be a fixed object point in the original coordinate system. Its coordinates $\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ in the rotated coordinate system can be calculated as

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = ROT^{-1}(\varphi, \lambda) \cdot \vec{o} \quad (\text{B.5})$$

with

$$ROT^{-1}(\varphi, \lambda) = \begin{pmatrix} \cos(\varphi) & 0 & -\sin(\varphi) \\ -\sin(\varphi)\sin(\lambda) & \cos(\lambda) & -\cos(\varphi)\sin(\lambda) \\ \sin(\varphi)\cos(\lambda) & \sin(\lambda) & \cos(\varphi)\cos(\lambda) \end{pmatrix}. \quad (\text{B.6})$$

x and y can be interpreted as coordinates of point \vec{o} in the projection plane of view (φ, λ) . $ROT_i = ROT(\varphi_i, \lambda_i)$, $i = 1, 2, 3$, is set for the given sample views and $\widehat{ROT} = ROT(\hat{\varphi}, \hat{\lambda})$ for the unfamiliar view $(\hat{\varphi}, \hat{\lambda})$, which is to be reconstructed. It is possible to number the sample views in such a way that

$$(iii) \quad \varphi_1 = \varphi_3 \quad \text{and}$$

$$(iv) \quad \lambda_1 = \lambda_2$$

hold. If the center view of the view bubble is the view with index 1 then the second and third views can be chosen in accordance with the conditions (i) and (ii). Furthermore,

$$ROT_i^{-1} = \begin{pmatrix} ROT_{iRow1}^{-1} \\ ROT_{iRow2}^{-1} \\ ROT_{iRow3}^{-1} \end{pmatrix} \quad (\text{B.7})$$

with ROT_{iRowj}^{-1} the j -th row vector of ROT_i^{-1} and

$$\widehat{ROT}^{-1} = \begin{pmatrix} \widehat{ROT}_{Row1}^{-1} \\ \widehat{ROT}_{Row2}^{-1} \\ \widehat{ROT}_{Row3}^{-1} \end{pmatrix}. \quad (\text{B.8})$$

Then following equations hold:

$$\begin{aligned} x_i &= ROT_{iRow1}^{-1} \cdot \vec{o} & y_i &= ROT_{iRow2}^{-1} \cdot \vec{o} \\ \hat{x} &= \widehat{ROT}_{Row1}^{-1} \cdot \vec{o} & \hat{y} &= \widehat{ROT}_{Row2}^{-1} \cdot \vec{o}. \end{aligned} \quad (\text{B.9})$$

B.1 Two Sample Views, x-Coordinate

Assumption (iii) and the equations (B.9) imply that \hat{x} can be expressed as linear combination of x_1 and x_2 , if $\widehat{ROT}_{Row1}^{-1}$ can be laid down as linear combination of ROT_{1Row1}^{-1} and ROT_{2Row1}^{-1} .

$$\widehat{ROT}_{Row1}^{-1} = a_1 \cdot ROT_{1Row1}^{-1} + a_2 \cdot ROT_{2Row1}^{-1} \quad (\text{B.10})$$

holds for $\det \begin{pmatrix} \cos(\varphi_1) & \cos(\varphi_2) \\ -\sin(\varphi_1) & -\sin(\varphi_2) \end{pmatrix} \neq 0$. This is equivalent to

$$\sin(\varphi_1 - \varphi_2) \neq 0. \quad (\text{B.11})$$

The last inequality holds for $\varphi_1 - \varphi_2 \neq k\pi$, $k \in \mathbb{Z}$. Due to assumption (i) that implies

$$\hat{x} = \sum_{i=1}^2 a_i x_i \quad (\text{B.12})$$

if

$$(b1 \ i) \quad \varphi_1 - \varphi_2 \neq \pm \pi \quad \text{and}$$

$$(b1 \ ii) \quad \varphi_1 \neq \varphi_2$$

and the coefficients a_i are derived from equation B.10.

B.2 Two Sample Views, y-Coordinate

The equations (B.9) imply that \hat{y} can be expressed as linear combination of x_1 , x_2 , and y_1 , if $\widehat{ROT}_{Row2}^{-1}$ can be expressed as linear combination of ROT_{1Row1}^{-1} , ROT_{2Row1}^{-1} , and ROT_{1Row2}^{-1} .

$$\widehat{ROT}_{Row2}^{-1} = b_1 \cdot ROT_{1Row1}^{-1} + b_2 \cdot ROT_{2Row1}^{-1} + b_3 \cdot ROT_{1Row2}^{-1} \quad (\text{B.13})$$

holds for $\det(ROT_{1Row1}^{-1}, ROT_{2Row1}^{-1}, ROT_{1Row2}^{-1}) \neq 0$. Due to assumption (iii) this is equivalent to

$$\cos(\lambda_1) \cdot \sin(\varphi_1 - \varphi_2) \neq 0. \quad (\text{B.14})$$

The last inequality holds if $\lambda_1 \neq k\frac{\pi}{2}$, $k \in \mathbb{Z}$, k odd, and $\varphi_1 - \varphi_2 \neq k'\pi$, $k' \in \mathbb{Z}$. Due to the assumptions (i) and (ii) that implies

$$\hat{y} = \sum_{i=1}^2 b_i \cdot x_i + b_3 \cdot y_1 \quad (\text{B.15})$$

if **(b1 i)**, **(b1 ii)**, and

$$\text{(b2iii)} \quad \lambda_1 \neq \frac{\pi}{2}$$

and the coefficients b_i are derived from equation B.13, e.g., by applying Cramer's rule.

B.3 Three Sample Views, x-Coordinate

Due to assumption (iii) and the zero entry in $ROT_{Row1}^{-1}(\varphi, \lambda)$, the formula for three sample views reduces to the same linear combination as for two sample views:

$$\hat{x} = \sum_{i=1}^2 a_i x_i \quad (\text{B.16})$$

if **(b1 i)** and **(b1 ii)** with the same coefficients a_i as in case B.1.

B.4 Three Sample Views, y-Coordinate

The equations (B.9) imply that \hat{y} can be expressed as linear combination of y_1 , y_2 , and y_3 , if $\widehat{ROT}_{Row2}^{-1}$ can be written as linear combination of ROT_{1Row2}^{-1} , ROT_{2Row2}^{-1} , and ROT_{3Row2}^{-1} .

$$\widehat{ROT}_{Row2}^{-1} = b_1 \cdot ROT_{1Row2}^{-1} + b_2 \cdot ROT_{2Row2}^{-1} + b_3 \cdot ROT_{3Row2}^{-1} \quad (\text{B.17})$$

holds for $\det(ROT_{1Row2}^{-1}, ROT_{2Row2}^{-1}, ROT_{3Row2}^{-1}) \neq 0$. Due to the assumptions (iii) and (iv) this is equivalent to

$$\sin(\lambda_1) \cdot \sin(\lambda_1 - \lambda_3) \cdot \sin(\varphi_1 - \varphi_2) \neq 0. \quad (\text{B.18})$$

The last inequality holds if $\lambda_1 \neq k\pi$, $\lambda_1 - \lambda_3 \neq k'\pi$, and $\varphi_1 - \varphi_2 \neq k''\pi$, $k, k', k'' \in \mathbb{Z}$. Due to the assumptions (i) and (ii) that implies

$$\hat{y} = \sum_{i=1}^3 b_i y_i \quad (\text{B.19})$$

if **(b1 i)**, **(b1 ii)**,

$$\text{(b4iii)} \quad \lambda_1 \neq \lambda_3 \quad \text{and}$$

$$\text{(b4iv)} \quad \lambda_1 \neq 0$$

and the coefficients b_i are derived from equation B.17.

List of Figures

1.1	Paintings from Monet’s Series “Rouen Cathedral”	2
1.2	Non-Valid Objects	3
1.3	Valid Objects	4
2.1	Aspect Graph	7
2.2	Canonical Views	8
3.1	Preprocessing	16
3.2	Self-Occlusions - Difference Between Object “Tom” and Object “Dwarf” . .	17
3.3	Robot Scene	18
3.4	Viewing Hemisphere	19
3.5	Segmentation Model	20
3.6	Shape of a Gabor Wavelet	21
3.7	Jet and Labeled Grid Graph	22
3.8	Matching Local Object Features	24
3.9	Tracking Local Object Features	25
4.1	View Bubble and its Approximation	28
4.2	Object “Tom”, Area of View Bubble	31
4.3	Object “Dwarf”, Area of View Bubbles	32
4.4	Qualitative Similarity Diagram	34
4.5	Object “Tom”, First Sequence With Similarity Diagram	35
4.6	Object “Dwarf”, First Sequence With Similarity Diagram	36
4.7	Canonical and Non-Canonical Views for Object “Tom”	37
5.1	Overlap of View Bubbles	42
5.2	Five Different Covers for Object “Tom”	44
5.3	Five Different Covers for Object “Dwarf”	45
5.4	Different Covers for Both Objects	46
5.5	Correlation between View Bubble Number and Similarity Threshold	47
5.6	Sparse Representation of Object “Tom”	48
6.1	Flowchart of View Morphing	50
6.2	Morphing from Three Sample Views	51
6.3	Longitude and Latitude Angles	52
6.4	Triangulation	54
6.5	Corresponding triangles in I and \hat{I}	55

6.6	Example of Morphed View for Object “Tom”	56
6.7	Example of Morphed View for Object “Dwarf”	57
6.8	Positions of Unfamiliar Views Which are Morphed for Statistics	58
6.9	Performance of the Relative Error	62
6.10	Correlation Between Morphing Errors and Number of View Bubbles	63
6.11	Correlation between Morphing Errors and Similarity Threshold	64
7.1	Flowchart of the Generation of Virtual Views	66
7.2	Weighting of Feature Vectors for Three Sample Views	67
7.3	Example of Virtual View Generation	71
7.4	Correlation Between Virtual View Errors and Number of View Bubbles	72
7.5	Correlation between Virtual View Errors and Similarity Threshold	73
8.1	Single Pose Estimation for Non-Degraded Images	77
8.2	Example for Single Pose Estimation	79
8.3	Single Pose and Sequence Estimation for Non-Degraded Images	83
8.4	Graphs Used for Noisy Test Images	84
8.5	Reconstructions of Non-Degraded and Noisy Images from Gabor Responses	87
8.6	Single Pose and Sequence Estimation for Noisy Images	88
8.7	Estimation Example - Object “Dwarf”, $\tau = 0.8$	89
8.8	Estimation Example - Object “Dwarf”, $\tau = 0.85$	90
8.9	Estimation Example - Object “Tom”, $\tau = 0.95$	91
A.1	Object “Tom”, Second Sequence With Similarity Diagram	96
A.2	Object “Tom”, Third Sequence With Similarity Diagram	97
A.3	Object “Tom”, Fourth Sequence With Similarity Diagram	98
A.4	Object “Dwarf”, Second Sequence With Similarity Diagram	99
A.5	Object “Dwarf”, Third Sequence With Similarity Diagram	100
A.6	Object “Dwarf”, Fourth Sequence With Similarity Diagram	101
A.7	Object “Tom”, Complete First Sequence	102
A.8	Object “Tom”, Complete Second Sequence	103
A.9	Object “Tom”, Complete Third Sequence	104
A.10	Object “Tom”, Complete Fourth Sequence	105
A.11	Object “Dwarf”, Complete First Sequence	106
A.12	Object “Dwarf”, Complete Second Sequence	107
A.13	Object “Dwarf”, Complete Third Sequence	108
A.14	Object “Dwarf”, Complete Fourth Sequence	109

List of Tables

4.1	Statistics for Object “Tom”	33
4.2	Statistics for Object “Dwarf”	33
5.1	Average Angle Between Center and East View	44
8.1	Mean Estimation Errors for Non-Degraded Images	78
8.2	Mean Estimation Errors for Degraded Images	85

Bibliography

- [1] D. J. Bartram. The Role of Visual and Semantic Codes in Object Naming. *Cognitive Psychology*, 6:325–356, 1974.
- [2] M. Becker, E. Kefalea, E. Maël, C. von der Malsburg, M. Pagel, J. Triesch, J. C. Vorbrüggen, and S. Zadel. GripSee: A Robot for Visually-Guided Grasping. In *Proceedings of ICANN International Conference on Artificial Neural Networks*, Skövde, Sweden, September 1998.
- [3] T. Beier and S. Neely. Feature-Based Image Metamorphosis. In *Proceedings of SIGGRAPH'92*, pages 35–42, 1992.
- [4] D. Beymer and T. Poggio. Face Recognition from One Example View. Technical Report CBCL Paper 121/AI Memo 1536, Massachusetts Institute of Technology, Cambridge, MA, September 1995.
- [5] I. Biederman. Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94:115–147, 1987.
- [6] V. Blanz, B. Schölkopf, H. H. Bülthoff, C. Burges, V. Vapnik, and T. Vetter. Comparison of View-Based Object Recognition Algorithms Using Realistic 3D Models. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff, editors, *Artificial Neural Networks - ICANN'96*, volume 1112, pages 251–256, Berlin, 1996. Springer Lecture Notes in Computer Science.
- [7] V. Blanz, M. J. Tarr, and H. H. Bülthoff. What Object Attributes Determine Canonical Views? *Perception*, 28:575–599, 1999.
- [8] M. C. A. Booth and E. T. Rolls. View-Invariant Representations of Familiar Objects by Neurons in the Inferior Temporal Visual Cortex. *Cerebral Cortex*, 8(6):510–523, 1998.
- [9] H. H. Bülthoff and S. Edelman. Psychophysical Support for a Two-Dimensional View Interpolation Theory of Object Recognition. In *Proceedings of the National Academy of Science of the United States of America*, volume 89, pages 60–64, 1992.
- [10] H. H. Bülthoff, S. Edelman, and M. J. Tarr. How are Three-Dimensional Objects Represented in the Brain? *Cerebral Cortex*, 5:247–260, 1995.
- [11] D. C. Burr, M. C. Morrone, and D. Spinelli. Evidence for Edge and Bar Detectors in Human Vision. *Vision Research*, 29(4):419–431, 1989.

- [12] C. G. Christou, B. S. Tjan, and H. H. Bülthoff. Old Paperclips, New Context. In *Investigative Ophthalmology & Visual Science, ARVO Annual Meeting*, volume 39(4), page 853, Fort Lauderdale, Florida, USA, May 10–15, 1998.
- [13] V. Chvatal. A Greedy Heuristic for the Set-Covering Problem. *Mathematics of Operations Research*, 4(3):233–235, 1979.
- [14] F. Cutzu and S. Edelman. Canonical Views in Object Representation and Recognition. *Vision Research*, 34:3037–3056, 1994.
- [15] C. Eckes and J. C. Vorbrüggen. Combining Data-Driven and Model-Based Cues for Segmentation of Video Sequences. In *Proceedings WCNN96*, pages 868–875, San Diego, CA, USA, 16–18 September, 1996. INNS Press & Lawrence Erlbaum Ass.
- [16] S. Edelman and H. H. Bülthoff. Orientation Dependence in the Recognition of Familiar and Novel Views of Three-Dimensional Objects. *Vision Research*, 32(12):2385–2400, 1992.
- [17] S. Edelman and D. Weinshall. A Selforganizing Multiple-View Representation of 3D Objects. *Biological Cybernetics*, 64(12):209–219, 1991.
- [18] D. J. Fleet and A. D. Jepson. Computation of Component Image Velocity from Local Phase Information. *International Journal of Computer Vision*, 5(1):77, 1990.
- [19] A. R. J. Francois and G. G. Medioni. Interactive 3D Model Extraction from a Single Image. *Image and Vision Computing*, 19(6):317–328, 2001.
- [20] M. Gray. Recognition Planning from Solid Models. In *Proceedings of the Alvey Computer Vision and Image Interpretation Meeting, Bristol*, pages 41–43, Sheffield, England, September 1986. Sheffield University Press.
- [21] K. L. Harman and G. K. Humphrey. Encoding “Regular” and “Random” Sequences of Views of Novel 3D Objects Rotating in Depth. In *Investigative Ophthalmology & Visual Science, ARVO Annual Meeting*, volume 39(4), page 856, Fort Lauderdale, Florida, USA, May 10–15, 1998.
- [22] J. K. Hietanen, D. I. Perrett, M. W. Oram, P. J. Benson, and W. H. Dittrich. The Effects of Lighting Conditions on the Responses of Cells Selective for Face Views in the Macaque Temporal Cortex. *Experimental Brain Research*, 89:157–171, 1992.
- [23] D. R. Hofstadter. *Gödel, Escher, Bach: an Eternal Golden Braid*. Basic Books, 1979.
- [24] J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987.
- [25] P. J. Kellman. Perception of Three-Dimensional Form in Infancy. *Perception and Psychophysics*, 36:353–358, 1984.
- [26] C. W. Khoh and P. Kovesi. *Animating Impossible Objects*. Department of Computer Science, The University of Western Australia, Nedlands, W.A. 6907, February 1999.

- [27] J. J. Koenderink and A. J. van Doorn. The Singularities of the Visual Mapping. *Biological Cybernetics*, 24:51–59, 1976.
- [28] J. J. Koenderink and A. J. van Doorn. The Internal Representation of Solid Shape with Respect to Vision. *Biological Cybernetics*, 32:211–216, 1979.
- [29] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen. Distortion Invariant Object Recognition in the Dynamic Link Architecture. *IEEE Transactions on Computers*, 42:300–311, 1993.
- [30] N. K. Logothetis, J. Pauls, H. H. Bülthoff, and Poggio T. Evidence for Recognition based on Interpolation among 2D Views of Objects in Monkeys. *Invest. Ophthalmol. Vis. Sci. Suppl.*, 34:1132, 1992.
- [31] N. K. Logothetis, J. Pauls, H. H. Bülthoff, and Poggio T. View-Dependent Object Recognition by Monkeys. *Current Biology*, 4:401–414, 1994.
- [32] N. K. Logothetis, J. Pauls, and Poggio T. Shape Representation in the Inferior Temporal Cortex of Monkeys. *Current Biology*, 5(5):552–563, 1995.
- [33] D. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Boston, MA, 1985.
- [34] R. Malik and T. Whangbo. Angle Densities and Recognition of 3D Objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):52–57, 1997.
- [35] D. Marr and H. K. Nishihara. Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes. In *Proceedings of the Royal Society of London, B(200)*, pages 269–294, 1978.
- [36] T. Maurer and C. von der Malsburg. Tracking and Learning Graphs and Pose on Image Sequences of Faces. In *Proceedings of the 2nd International Conference on Automatic Face- and Gesture- Recognition*, pages 176–181, Killington, Vermont, USA, October 1996.
- [37] K. Mehlhorn and S. Näher. Dynamic Delaunay Triangulations. 1998.
- [38] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Introduction of the Metropolis Algorithm for Molecular-Dynamics Simulation. *J. Chem. Phys.*, 21:1987, 1953.
- [39] Y. Miyashita. Associative Representation of Visual Long Term Memory in the Neurons of the Primate Temporal Cortex. In E. Iwai and M. Mishkin, editors, *Vision, Memory and the Temporal Lobe*, pages 75–87. Elsevier, New York, 1990.
- [40] J. J. More. The Levenberg-Marquardt Algorithm: Implementation and Theory. In G. A. Watson, editor, *Lecture Notes in Mathematics - Numerical Analysis*, number 630, pages 105–116. Springer, 1978.
- [41] H. Murase and S. Nayar. Visual Learning and Recognition of 3-D Objects from Appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

- [42] R. Nevatia and T. O. Binford. Structured Descriptions of Complex Objects. In Nils J. Nilsson, editor, *Proceedings of the 3rd International Joint Conference on Artificial Intelligence*, pages 641–647, Standford, CA, 1973.
- [43] R. Nevatia and T. O. Binford. Description and Recognition of Curved Objects. *Artificial Intelligence*, 8(1):77–98, 1977.
- [44] T. Niemann, M. Lappe, and K.-P. Hoffmann. Visual Inspection of Three-Dimensional Objects by Human Observers. *Perception*, 25:1027–1042, 1996.
- [45] S. E. Palmer, E. Rosch, and P. Chase. Canonical Perspective and the Perception of Objects. In I. Long and A. Baddeley, editors, *Attention and Performance IX*, pages 135–151. Erlbaum, Hillsdale, N.J., 1981.
- [46] D. I. Perrett and M. H. Harries. Characteristic Views and the Visual Inspection of Simple Faceted and Smooth Objects: “Tetraheder and Potatoes”. *Perception*, 17:703–720, 1988.
- [47] D. I. Perrett, A. J. Mistlin, and A. J. Chitty. Visual Neurons Responsive to Faces. *Trends in Neurosciences*, 10:358–364, 1989.
- [48] D. I. Perrett, M. W. Oram, M. H. Harries, Bevan R., J. K. Hietanen, P. J. Benson, and S. Thomas. Viewer-Centred and Object-Centred Coding of Heads in the Macaque Temporal Cortex. *Experimental Brain Research*, 86:159–173, 1991.
- [49] D. I. Perrett, M. W. Oram, J. K. Hietanen, and P. J. Benson. Issues of Representation in Object Vision. In M. J. Farah and G. Ratcliff, editors, *The Neuropsychology of High-Level Vision - Collected Tutorial Essays*, pages 33–61, Hillsdale, New Jersey, 1994. Lawrence Erlbaum Associates.
- [50] D. I. Perrett, P. A. J. Smith, D. D. Potter, A. J. Mistlin, A. S. Head, A. D. Milner, and M. A. Jeeves. Visual Cells in the Temporal Cortex Sensitive to Face View and Gaze Direction. In *Proceedings of the Royal Society of London, B(3)*, pages 293–317, 1985.
- [51] G. Peters. Theories of Three-Dimensional Object Perception - A Survey. In *Recent Research Developments in Pattern Recognition, Part-I*, volume 1, pages 179–197. Transworld Research Network, 2000.
- [52] G. Peters and C. von der Malsburg. Interpolation of Novel Object Views from Sample Views. In Milad Fares Sebaaly, editor, *Proceedings of the International NAISO Congress on Information Science Innovations (ISI'2001)*, pages 800–805, Dubai, U.A.E., March 17 - 21, 2001.
- [53] G. Peters and C. von der Malsburg. View Reconstruction by Linear Combination of Sample Views. In T. Cootes and C. Taylor, editors, *Proceedings of the British Machine Vision Conference 2001 (BMVC:2001)*, pages 223–232, Manchester, UK, September 10-13, 2001.

- [54] G. Peters, B. Zitova, and C. von der Malsburg. A Comparative Evaluation of Matching and Tracking Object Features for the Purpose of Estimating Similar-View-Areas of 3-Dimensional Objects. Technical Report IRINI 99-06, Institut für Neuroinformatik, Ruhr-Universität Bochum, D-44780 Bochum, Germany, April 1999.
- [55] G. Peters, B. Zitova, and C. von der Malsburg. Two Methods for Comparing Different Views of the Same Object. In T. Pridmore and D. Elliman, editors, *Proceedings of the 10th British Machine Vision Conference (BMVC'99)*, pages 493–502, Nottingham, UK, September 13-16, 1999.
- [56] T. Poggio and S. Edelman. A Network that Learns to Recognize Three-Dimensional Objects. *Nature*, 343:263–266, 1990.
- [57] M. Pöttsch. Die Behandlung der Wavelet-Transformation von Bildern in der Nähe von Objektkanten. Diploma Thesis, Technical Report IRINI 94-04, Institut für Neuroinformatik, Ruhr-Universität Bochum, D-44780 Bochum, Germany, Mai 1994.
- [58] M. Pöttsch, T. Maurer, L. Wiskott, and C. von der Malsburg. Reconstruction from Graphs Labeled with Responses of Gabor Filters. In C. von der Malsburg, W. von Seelen, J. Vorbrüggen, and B. Sendhoff, editors, *Proceedings of the ICANN 96*, pages 851–856, Bochum, Germany, July, 1996.
- [59] I. Rock and J. DiVita. A Case of Viewer-Centered Object Perception. *Cognitive Psychology*, 19:280–293, 1987.
- [60] S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290:2323–2326, 2000.
- [61] B. Schölkopf. *Support Vector Learning*. Ph. D. Thesis, Informatik der Technischen Universität Berlin, 1997.
- [62] M. Seibert and A. M. Waxman. Adaptive 3-D Object Recognition from Multiple Views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):107–124, 1992.
- [63] S. M. Seitz and C. R. Dyer. Physically-Valid View Synthesis by Image Interpolation. In *Proceedings of the Workshop on Representations of Visual Scenes*, Cambridge, MA, 1995.
- [64] S. M. Seitz and C. R. Dyer. Toward Image-Based Scene Representation Using View Morphing. In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 1, pages 84–89, Vienna, Austria, 1996.
- [65] S. M. Seitz and C. R. Dyer. View Morphing. In *Proceedings of SIGGRAPH'96*, pages 21–30, 1996.
- [66] H. S. Seung and D. D. Lee. The Manifold Ways of Perception. *Science*, 290:2268–2269, 2000.
- [67] R. N. Shepard and J. Metzler. Mental Rotation of Three-Dimensional Objects. *Science*, 171:701–703, 1971.

- [68] M. J. Tarr. *Orientation Dependence in Three-Dimensional Object Recognition*. Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1989.
- [69] M. J. Tarr and S. Pinker. When does Human Object Recognition use a Viewer-Centered Reference Frame? *Psychological Science*, 1:253–256, 1990.
- [70] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290:2319–2322, 2000.
- [71] W. M. Theimer and H. A. Mallot. Phase-Based Binocular Vergence Control and Depth Reconstruction using Active Vision. *CVGIP: Image Understanding*, 60(3):343, 1994.
- [72] J. T. Todd and F. D. Reichel. Ordinal Structure in the Visual Perception and Cognition of Smoothly Curved Surfaces. *Psychol. Rev.*, 96:643–657, 1989.
- [73] S. Ullman. Aligning Pictorial Descriptions: An Approach to Object Recognition. *Cognition*, 32:193–254, 1989.
- [74] S. Ullman and R. Basri. Recognition by Linear Combinations of Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):992–1006, 1991.
- [75] Jan C. Vorbrüggen. *Zwei Modelle zur datengetriebenen Segmentierung visueller Daten*, volume 47 of *Reihe Physik*. Verlag Harri Deutsch, Thun, Frankfurt am Main, Germany, 1995.
- [76] E. K. Warrington and A. M. Taylor. The Contribution of the Right Parietal Lobe to Object Recognition. *Cortex*, 9:152–164, 1973.
- [77] D. Weinshall and M. Werman. On View Likelihood and Stability. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):97–108, 1997.
- [78] E. W. Weisstein. *CRC Concise Encyclopedia of Mathematics*. Chapman & Hall/CRC, Boca Raton, 1999.
- [79] M. Wertheimer. Untersuchung zur Lehre von der Gestalt II. *Psychologische Forschung*, 4:301–350, 1923.
- [80] M. Wexler, S. M. Kosslyn, and A. Berthoz. Motor processes in mental rotation. *Cognition*, 68:77–94, 1998.
- [81] J. Wieghardt. *Learning the Topology of Views: From Images to Objects*. Ph. D. Thesis, Institut für Neuroinformatik, Ruhr-Universität Bochum, 2001.
- [82] L. Wiskott. *Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis*, volume 53 of *Reihe Physik*. Verlag Harri Deutsch, Thun, Frankfurt am Main, Germany, 1995.
- [83] G. Wolberg. *Digital Image Warping*. IEEE Computer Society Press, Los Alamitos, CA, 1990.

Previously Published Contents of this Thesis

Parts of chapter 2 have already been published in

Gabriele Peters, *Theories of Three-Dimensional Object Perception - A Survey*, Recent Research Developments in Pattern Recognition, Part-I, vol. 1, pp. 179–197, Transworld Research Network, Kerala, Trivandrum-8, India, 2000.

Parts of chapters 3 and 4 have already been published in

Gabriele Peters, Barbara Zitova, and Christoph von der Malsburg, *Two Methods for Comparing Different Views of the Same Object*, Proceedings of the 10th British Machine Vision Conference (BMVC'99), Tony Pridmore and Dave Elliman (editors), University of Nottingham, vol. 2, pp. 493-502, Nottingham, UK, September 13-16, 1999

and

Gabriele Peters, Barbara Zitova and Christoph von der Malsburg, *A Comparative Evaluation of Matching and Tracking Object Features for the Purpose of Estimating Similar-View-Areas of 3-Dimensional Objects*, Technical Report, IRINI 99-06, Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany, April, 1999.

Parts of chapters 3 and 7 have already been published in

Gabriele Peters and Christoph von der Malsburg, *Interpolation of Novel Object Views from Sample Views*, Proceedings of the International NAISO Congress on Information Science Innovations (ISI'2001), Milad Fares Sebaaly (editor), pp. 800-805, Dubai, U.A.E., March 17 - 21, 2001

and

Gabriele Peters, *The Reconstruction of Unfamiliar Views of a 3-D Object from a Sparse Set of Familiar Views*, Conference on Cognitive Neuroscience, Abstract Book, Hanse Wissenschaftskolleg, Bremen, Germany, October 31 - November 3, 1999.

Parts of chapters 3, 5, 6, and 7 have already been published in

Gabriele Peters and Christoph von der Malsburg, *View Reconstruction by Linear Combination of Sample Views*, Proceedings of the 12th British Machine Vision Conference (BMVC:2001), Tim Cootes and Chris Taylor, University of Manchester, vol. 1, pp. 223-232, Manchester, UK, September 10-13, 2001.

Parts of chapter 7 have already been published in

Christoph von der Malsburg, Kurt Reiser, Gabriele Peters, Jan Wiegardt, and Kazunori Okada, *3D Object Representation by 2D Views*, Proceedings of the 6th ATR Symposium on Face and Object Recognition, pp. 11-12, Kyoto, Japan, July, 1999.