Qualitative Multi-Scale Feature Hierarchies for Object Tracking*

Lars Bretzner and Tony Lindeberg

Computational Vision and Active Perception Laboratory (CVAP)
Department of Numerical Analysis and Computing Science
KTH (Royal Institute of Technology)
S-100 44 Stockholm, Sweden.

Email: { bretzner, tony}@nada.kth.se

Technical report ISRN KTH/NA/P/–9909–SE

Abstract

This paper shows how the performance of feature trackers can be improved by building a hierarchical view-based object representation consisting of qualitative relations between image structures at different scales. The idea is to track all image features individually, and to use the qualitative feature relations for avoiding mismatches, resolving ambiguous matches and for introducing feature hypotheses whenever image features are lost. Compared to more traditional work on view-based object tracking, this methodology has the ability to handle semi-rigid objects and partial occlusions. Compared to trackers based on three-dimensional object models, this approach is much simpler and of a more generic nature. A hands-on example is presented showing how an integrated application system can be constructed from conceptually very simple operations.

*The support from the Swedish Research Council for Engineering Sciences, TFR, and the Swedish National Board for Industrial and Technical Development, NUTEK, is gratefully acknowledged. Accepted for publication in Journal of Visual Communication and Image Representation. An earlier version of this manuscript was presented in M. Nielsen, P. Johansen, O. Olsen and J. Weickert (eds), *Proc. Second International Conference on Scale-Space Theories in Computer Vision*, (Corfu, Greece), September 1999. Springer-Verlag Lecture Notes in Computer Science, vol 1682, pp. 117–128.

1 Introduction

To maintain a stable representation of a dynamic world, it is necessary to relate image data from different time moments. When analysing image sequences frame by frame, as is commonly done in computer vision applications, it is therefore useful to include an explicit tracking mechanisms into the vision system.

When constructing such a tracking mechanism, there is a large freedom in design, concerning how much a priori information should be included into and be used by the tracker. If the goal is to track a single object of known shape, then it may be natural to build a three-dimensional object model, and to relate computed views of this internal model to the image data that occur. An alternative approach is store a large number of actual views in a database, and subsequently match these to the image sequence.

Depending on what type of object representation we choose, we can expect different trade-offs between the complexity of constructing the object representation and the complexity in matching the object representation to image data.[1] In particular, different design strategies will imply different amounts of additional work when the database is extended with new objects.

The subject of this article is to advocate the use of *qualitative* multi-scale object models in this context, as opposed to more detailed models. The idea is to represent only dominant image features of the object, and relations between those that are reasonably stable under view variations. In this way, a new object model can be constructed with only minor additional work, and it will be demonstrated that such a weaker approach to object representation is powerful enough to give a significant improvement in the robustness of feature trackers.

A main rationale for the proposed approach is that if we track individual features over long time periods in scenes with changing conditions (e.g., object pose and illumination), the likelihood that features will be mismatched or lost will increase with time. Major aims of the proposed hierarchical representation are to handle such problems, and also to assist in the initialization stage of the feature tracker. When a feature is lost, the relations of the qualitative feature hierarchy model will be used for defining search regions in the which the lost feature can be detected. When mismatches occur, relational constraints in the feature hierarchy will be helpful for detecting and rejecting outliers.

The usefulness of such a hierarchical object representation for feature tracking will be demonstrated by experiments on real-world image sequences. Specifically, it will be shown how an integrated non-trivial application to human-computer interaction can be constructed in a straightforward and conceptually very simple way, by combination with a set of elementary scale-space operations.

The presentation is organized as follows: Section 2 presents the general motivations behind the proposed approach, with an overview of related works. In section 3, we first briefly review the multi-scale framework we use for detecting image features, and describe how hierarchical and qualitative feature relations can be defined between these multi-scales image features. Section 4 outlines how such a view-based object

[1] With the term "complexity", we here refer to both the computational complexity in matching algorithms and the degree of structural complexity that is required when designing the software.

representation can be used in the context of feature tracking, and shows experimental results for two sample applications to hand gesture analysis and face tracking, respectively. Finally, section 5 concludes with a summary and discussion concerning other possible applications and generalizations of the proposed ideas.

2 Choice of Image Representation for Feature Tracking

The framework we consider is one in which image features are detected at multiple scales. Each feature is associated with a region in space as well as a range of scales, and relations between features at different scales impose hierarchical links across scales. Specifically, we assume that the image features are detected with a mechanism for automatic scale selection (Lindeberg 1998*b*). In earlier work (Bretzner & Lindeberg 1998*a*), we have demonstrated how such a scale selection mechanism is essential to obtain a robust behaviour of the feature tracker if the image features undergo large size variations in the image domain.

The rationale for using a hierarchical multi-scale image representation for feature tracking originates from the well-known fact that real-world objects consist of different types of structures at different scales. An internal object representation should reflect this fact. One aspect of this, which we shall make particular use of, is that certain hierarchical relations over scales tend to remain reasonably stable when the viewing conditions are varied. Thus, even if some features are lost during tracking (e.g. due to occlusions, illumination variations, or spurious errors by the feature detector or the feature matching algorithm), it is rather likely that a sufficient number of image features will remain to support the tracking of the other features. Thereby, the feature tracker will have higher robustness[2] with respect to occlusions, viewing variations and spurious errors in the lower-level modules. As we shall see, the qualitative nature of these feature relations will also make it possible to handle semi-rigid objects within the same framework.

In this way, the approach we will propose is closely related to the notion of object representation. Compared to the more traditional problem of object recognition, however, the requirements are different, since the primary goal is to maintain a stable image representation over time, and we do not need to support indexing and recognition functionalities into large databases. For these reasons, a qualitative image representation can be sufficient in many cases, and offer a higher flexibility by being more generic than detailed object models.

Related works. The topic of this paper touches on both the subjects of feature tracking and object representation. The literature on tracking is large and impossible to review here. Hence, we focus on the most closely related works.

Image representations involving linking across scales have been presented by several authors. (Crowley & Parker 1984, Crowley & Sanderson 1987) detected peaks and ridges in a pyramid representation. In retrospect, a main reason why stability problems were encountered is that the pyramids involved a rather coarse sampling in

[2]According to the terminology proposed by (Toyama & Hager 1999), the automatic scale selection mechanism is essential for the pre-failure robustness of the feature tracker, while the proposed qualitative multi-scale feature hierarchy improves the post-failure robustness.

the scale direction. (Koenderink 1984) defined links across scales using iso-intensity paths in scale-space, and this idea was made operational for medical image segmentation by (Lifshitz & Pizer 1990) and (Vincken et al. 1997). (Lindeberg 1993) constructed a scale-space primal sketch, in which a morphological support region was associated with each extremum point and paths of critical points over scales were computed delimited by bifurcations. (Olsen 1997) applied a similar approach to watershed minima in the gradient magnitude. (Griffin et al. 1992) developed a closely related approach based on maximum gradient paths, however, at a single scale. In the scale-space primal sketch, scale selection was performed, by maximizing measures of blob strength over scales, and significance was measured by the volumes that image structures occupy in scale-space, involving the stability over scales as a major component. A generalization of this scale selection idea to more general classes of image structures was presented in (Lindeberg 1994, Lindeberg 1998*b*, Lindeberg 1998*a*), by detecting scale-space maxima, *i.e.* points in scale-space at which normalized differential measures of feature strength assume local maxima with respect to scale. (Pizer et al. 1994) and his co-workers (Gauch & Pizer 1993) have proposed closely related descriptors, focusing on multi-scale ridge representations for medical image analysis. Psychophysical results by (Burbeck & Pizer 1995) support the belief that such hierarchical multi-scale representations are relevant for object representation.

With respect to the problem of object recognition, (Shokoufandeh et al. 1998) detect extrema in a wavelet transform in a way closely related to the detection of scale-space maxima, and define a graph structure from these image features. This graph structure is then matched to corresponding descriptors for other objects, based on topological and geometric similarity. Earlier graph-like object representations include the classical model-based approach by (Lowe 1985), used in conjunction with perceptual grouping, as well as the distributed aspect hierarchy proposed by (Dickinson et al. 1992). In relation to the large number of works on model based tracking, there are similar aims between our approach and the following works: (Koller et al. 1993) used car models to support the tracking of vehicles in long sequences with occlusions and illumination variations. (Smith & Brady 1995) defined clusters of coherently moving corner features as to support the tracking of cars in a qualitative manner. (Black & Jepson 1998*b*) constructed a view-based object representation using an eigenimage approach to compactly represent and support the tracking of an object seen from a large number of different views. The recently developed condensation algorithm (Isard & Blake 1998, Black & Jepson 1998*a*) is of particular interest, by explicitly constructing statistical distributions to capture relations between image features. Concerning the specific application to qualitative hand tracking that will be addressed in this paper, more detailed hand models have been presented by (Kuch & Huang 1995, Heap & Hogg 1996, Yasumuro et al. 1999). Related graph-like representations for hand tracking and face tracking have been presented by (Triesch & von der Malsburg 1996, Mauerer & von der Malsburg 1996).

3 Image Features and Qualitative Feature Relations

We are interested in representing objects which can give rise to a rich variety of image features of different types and at different scales. Generically, these image features

can be (i) zero-dimensional (junctions), (ii) one-dimensional (edges and ridges), or (iii) two-dimensional (blobs), and we assume that each image feature is associated with a region in space as well as a range of scales.

3.1 Computation of Image Features

When computing a hierarchical view-based object representation, one may at first desire to compute a detailed representation of the multi-scale image structure, as done by the scale-space primal sketch or some of the closely related representations reviewed in section 2. Since we are interested in processing temporal image data, however, and the construction of such a representation from image data requires a rather large amount of computations, we shall here follow a computationally more efficient approach.

We focus on image features expressed in terms of *scale-space maxima*, *i.e.* points in scale-space at which differential geometric entities assume local maxima with respect to space and scale (Lindeberg 1998*b*). Formally, such points are defined by

$$(\nabla (\mathcal{D}_{norm} L(x;\ s)) = 0) \quad \wedge \quad (\partial_s (\mathcal{D}_{norm} L(x;\ s)) = 0) \tag{1}$$

where $L(\cdot;\ s)$ denotes the scale-space representation of the image f constructed by convolution with a Gaussian kernel $g(\cdot; s)$ with scale parameter (variance) s and $\mathcal{D}_{norm}$ is a differential invariant normalized by the replacement of all spatial derivatives ∂_{x_i} by γ-normalized derivatives $\partial_{\xi_i} = s^{\gamma/2} \partial_{x_i}$.

Two examples of such differential descriptors, which we shall make particular use of here, include the normalized Laplacian (with $\gamma = 1$) for blob detection

$$\nabla^2_{norm} L = s\,(L_{xx} + L_{yy}) \tag{2}$$

and the square difference between the eigenvalues L_{pp} and L_{qq} of the Hessian matrix (with $\gamma = 3/4$) for ridge detection

$$\mathcal{A} L_{\gamma-norm} = s^{2\gamma}\,|L_{pp} - L_{qq}|^2 = s^{2\gamma}\,((L_{xx} - L_{yy})^2 + 4 L_{xy}^2) \tag{3}$$

see (Lindeberg 1998*a*) for a more general description. A computationally very attractive property of this construction is that the scale-space maxima can be computed by architecturally very simple and computationally highly efficient operations involving: (i) scale-space smoothing, (ii) pointwise computation of differential invariants, and (iii) detection of local maxima of scalar entities in scale-space.

Furthermore, to simplify the geometric analysis of image features, we shall reduce the spatial representation of image descriptors to ellipses, by evaluating a second moment matrix

$$\mu = \int_{\eta \in \mathbb{R}^2} \begin{pmatrix} L_x^2 & L_x L_y \\ L_x L_y & L_y^2 \end{pmatrix} g(\eta; s_{int})\, d\eta \tag{4}$$

at integration scale s_{int} proportional to the detection scale of the scale-space maximum (equation (1)). Thereby, each image feature will we represented by a point $(x;\ s)$ in scale-space and a covariance matrix Σ describing the shape, graphically illustrated by an ellipse. For one-dimensional features, the corresponding ellipses will be highly

elongated, while for zero-dimensional and two-dimensional features, the ellipse descriptors of the second moment matrices will be rather circular. Attributes derived from the covariance matrix include its anisotropy derived from the ratio $\lambda_{max}/\lambda_{min}$ between its eigenvalues, and its orientation defined as the orientation of its main eigenvector.

Figure 4 shows an example of such image descriptors computed from a grey-level image, after ranking on a significance measure defined as the magnitude of the response of the differential operator at the scale-space maximum. A trivial but nevertheless very useful effect of this ranking is that it substantially reduces the number of image features for further processing, thus improving the computational efficiency. In a more detailed representation of the multi-scale deep structure of a real-world image, it will often be the case that a large number of the image features and their hierarchical relations correspond to image structures that will be regarded as insignificant by later processing stages.

3.2 Qualitative Feature Relations

Between the abovementioned features, various types of relations can be defined in the image plane. Here, we consider the following types of qualitative relations:

Spatial coincidence (inclusion): We say that a region A at position x_A and scale s_A is in spatial coincidence relation to a region B at position x_B and at a (coarser) scale $s_B > s_A$ if

$$(x_A - x_B)^T \Sigma_B^{-1} (x_A - x_B) \in [D_1, D_2] \tag{5}$$

where D_1 and D_2 are distance thresholds and Σ_B is a covariance matrix associated with region B. By using a Mahalanobis distance measure, we introduce a directional preference which is highly useful for expressing spatial relations between elongated image features. While the special case $D_1 = 0$ corresponds to an inclusion relation, there are also cases where one may want to explicitly represent distant features, using $D_1 > 0$

Stability of scale relations: For two image features at times t_k and $t_{k'}$, we assume that the ratio between their scale values should be approximately the same. This is motivated by the physical requirement of scale invariance under zooming

$$\frac{s_A(t_k)}{s_B(t_k)} \approx \frac{s_A(t_{k'})}{s_B(t_{k'})}. \tag{6}$$

To accept small variations due to changes in view direction and spurious variations from the scale selection mechanism of the feature tracker, we measure relative distances in the scale direction and implement the "$\approx$" operation by $q \approx q' \iff |\log \frac{q}{q'}| < \log T$, where $T > 1$ is a threshold in the scale direction.

Directional relation (bearing): For a feature A related to a one-dimensional feature B, the angle is measured between the main eigenvector of Σ_B and the vector $x_A - x_B$ from the center x_B of B to the center x_A of A (see Figure 1) .

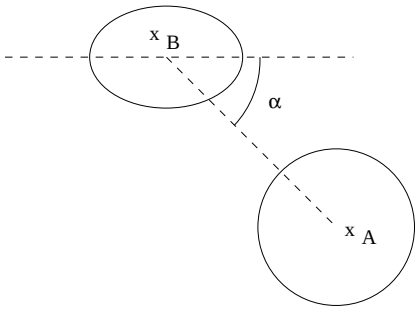

Figure 1: The direction relation (bearing) between two features A and B is the angle α between the main eigenvector of Σ_B (illustrated by the ellipse) and the vector $x_A - x_B$.

Trivially, these relations are invariant to translations and rotations in the image plane. The scale invariance of these relations follows from corresponding scale invariance properties of image descriptors computed from scale-space maxima — if the size of an image structure is scaled by a factor c in the image domain, then the corresponding scale levels are transformed by a factor c^2.

3.3 Qualitative Multi-Scale Feature Hierarchy

Let us now consider a specific example with images of a hand. From our knowledge that a hand consists of five fingers, we construct a model consisting of: (i) the palm, (ii) the five fingers, (iii) a finger tip for each finger, (see figure 2).

Each finger is in a spatial coincidence relation to the palm, as well as a directional relation. Moreover, each fingertip is in a spatial relationship to its finger, and satisfies a directional relation to this feature. In a similar manner, each finger is in a scale stability relation with respect to the palm, and each fingertip is in a corresponding scale stability relation relative to its finger.

Such a representation will be referred to as a *qualitative multi-scale feature hierarchy*. Figure 3 shows the relations this representation is built from, using UML notation (Fowler & Scott 1997). An attractive property of this view-based object representation is that it only focuses on qualitative object features. There is no assumption of rigidity, only that the qualitative shape is preserved.

The idea behind this construction is of course that the palm and the fingertips should give rise to blob responses (equation (2)) and that the fingers give rise to ridge responses (equation (3)). Figure 4 shows an example of how this model can be initialized and matched to image data with associated image descriptors.

To exclude responses from the background, we have here required that all image features should correspond to bright blobs or bright ridges. Alternatively, one could define spatial inclusion relations with respect to other segmentation cues relative to the background, *e.g.* chromaticity or depth.

Here, we have constructed the graph with feature relations manually, using qualitative knowledge about the shape of the object and its primitives. In a more general setting, however, one can also consider the learning of stable feature relations in an actual setting, based on a richer set of image features as well as a richer vocabulary of

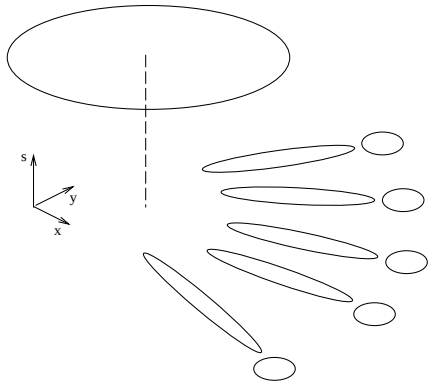

Figure 2: A qualitative multi-scale feature hierarchy constructed for a hand model.

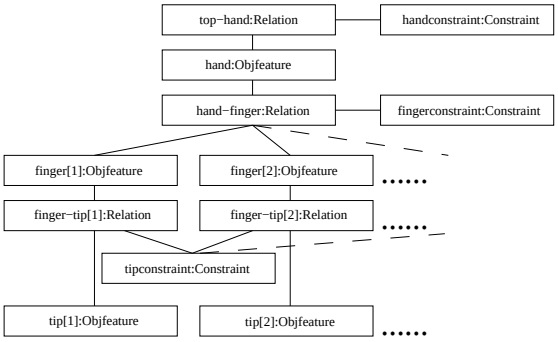

Figure 3: Instance diagram for the feature hierarchy of a hand (figure 2).

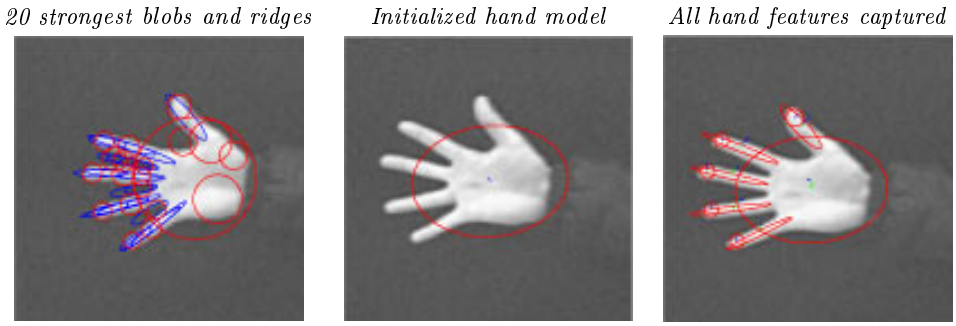

Figure 4: Illustration of the initialization stage of the object tracker. Once the coarse-scale feature is found (here the palm of the hand), the qualitative feature hierarchy guides the top-down search for the remaining features of the representation. (The left image shows the 20 most significant blob responses (in red) and ridge responses (in blue).)

qualitative feature relations. Of particular interest may be to learn probability distributions of the relations between image features, in a similar spirit as the condensation algorithm (Isard & Blake 1998, Black & Jepson 1998*a*).

Moreover, the list of feature relations in section 3.2 should by no means be regarded as exhaustive. Additional feature relations can be introduced whenever motivated by their effectiveness in specific applications. For example, in several cases it is natural to introduce a richer set of inter-feature relations between the primitives that are the ancestors of a coarser scale image feature.

4 Feature Tracking with Hierarchical Support

One idea that we are going to make explicit use of in this paper is to let features at different scales support each other during feature tracking. If fine-scale features are lost, then the coarse scale features combined with the other fine-scale features should provide sufficient information so as to generate hypotheses for recapturing the lost feature. Similarly, if a coarse scale feature is lost, *e.g.* due to occlusion or a too large three-dimensional rotation, then the fine-scale features should support the model based tracking. While this behaviour can be easily achieved with a three-dimensional object model, we are here interested in generic feature trackers which operate without detailed quantitative geometric information.

Figure 5 gives an overview of the composed object tracking scheme. The feature tracking module underlying this scheme is described in (Bretzner & Lindeberg 1998*a*), and consists of the evaluation of a multi-cue similarity measure involving patch correlation, and stability of scale descriptors and significance measures for image features detected according to section 3.1.

Scheme for object tracking using qualitative feature hierarchies:

Initialization:

Find and match top-level feature using initial position and top-level parent-children constraints.

Tracking:

For each frame:

For each feature in the hierarchy (top-down):

Track image features (see separate description)

If a feature is lost (or not found)

If parent matched

Find feature using parent position and parent-feature relation constraints

else if child(ren) matched

Find feature using child(ren) position and feature-children relation constraints.

Parse feature hierarchy, verify relations and reject mismatches.

Figure 5: Overview of the scheme for object tracking with hierarchical support.

4.1 Sample Application I — The 3-D Hand Mouse

From the a set of trajectories of image features extracted from an object, we can compute the motion of the object, assuming that a sufficient number of image features is available and that the object is kept rigid. One application that we are particularly interested in is to use such motions as mediated by hand gestures for controlling other computerized equipment (Lindeberg & Bretzner 1998). Examples of applications of this idea include:

- interaction with visualization systems and virtual environments,
- control of mechanical systems, and
- immaterial remote control functionality for consumer electronics.

Related works in this direction have been presented by (Cipolla et al. 1993, Freeman & Weissman 1995, Cipolla & Hollinghurst 1996, Maggioni & Kämmerer 1998).

The mathematical foundation underlying this "3-D hand mouse" was presented in (Bretzner & Lindeberg 1998*b*), in the form of a general framework for computing three-dimensional structure and motion from a set of sparse point and line features in multiple affine views. (Here, the point features correspond to blob responses from the finger tips, while the line features capture the orientations of ridge descriptors extracted from the fingers.) Our previous experimental work, however, was done with image sequences where an individual feature tracker with automatic scale selection (Bretzner & Lindeberg 1998*a*) was sufficient to obtain the extended feature trajectories needed for structure and motion computations.

The qualitative feature hierarchy provides a useful tool for extending this functionality, by making the system less sensitive to spurious errors when tracking image features individually. First of all, figure 6 demonstrates the ability of the qualitative feature hierarchy to handle non-rigid motions. Since the relations between the image features are of a qualitiative nature, it follows that these feature relations will remain valid under moderate perturbations of positions of the image features. Then, figures 7–8 show two examples of how this view-based object representation supports the recapturing of lost image features. In the first sequence, one finger is first lost due to occlusion and later recaptured. In the next sequence, the hand is turning and almost all features are lost except the top level feature (the blob corresponding to the palm). When the features are no longer occluded, the tracker captures them in a coarse-to-fine manner according to the scheme in figure 5.

While the object representation underlying these computations is a view-based representation, it should be remarked that the step is not far to a three-dimensional object model. If the hand is kept rigid over a sufficiently large three-dimensional rotation, we can use the motion information in the feature trajectories of the fingers and the finger tips for computing the structure and the motion of the object (see (Bretzner & Lindeberg 1998*b*) for algorithmic details). Figure 9 shows an illustration of this 3-D hand mouse in operation. The left column shows four different snapshots from a sequence of a moving hand, while the right column shows corresponding motion estimates visualized by subjecting a cube to the estimated three-dimensional rotation.

Figure 10 gives an overview of the components involved in our current implementation of this computer vision based interface for human-computer interaction. With

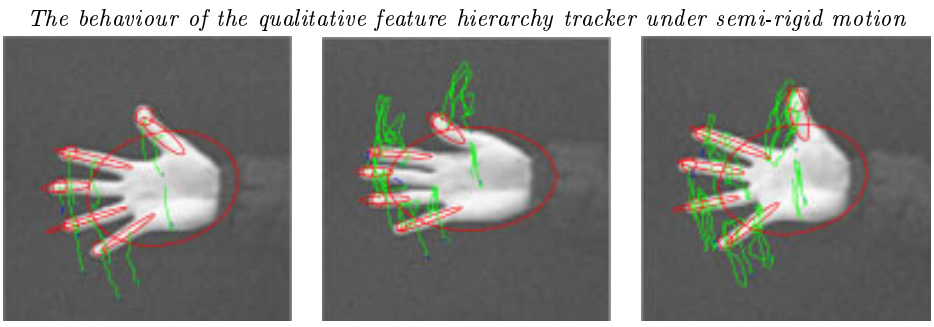

Figure 6: Due to the qualitative nature of the feature relations, the proposed framework allows objects to be tracked under semi-rigid motion.

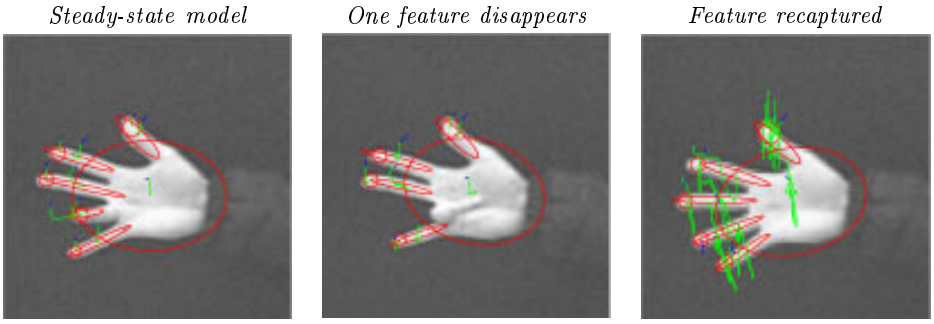

Figure 7: The proposed qualitative representation makes it possible to maintain tracking even if parts of the object are occluded. Later in the sequence, the occluded part (in this case the finger), can be captured again using the feature hierarchy. (Here, all image features are illustrated by red, while the feature trajectories are green.)

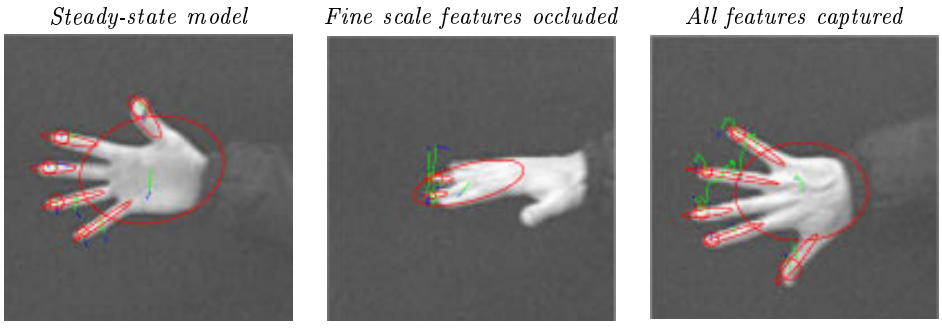

Figure 8: Illustration of how the qualitative feature hierarchy makes it possible to maintain object tracking under view variations. The images show how most finger features are lost due to occlusion when the hand turns, and how the qualitative feature hierarchy guides the search to find these features again.

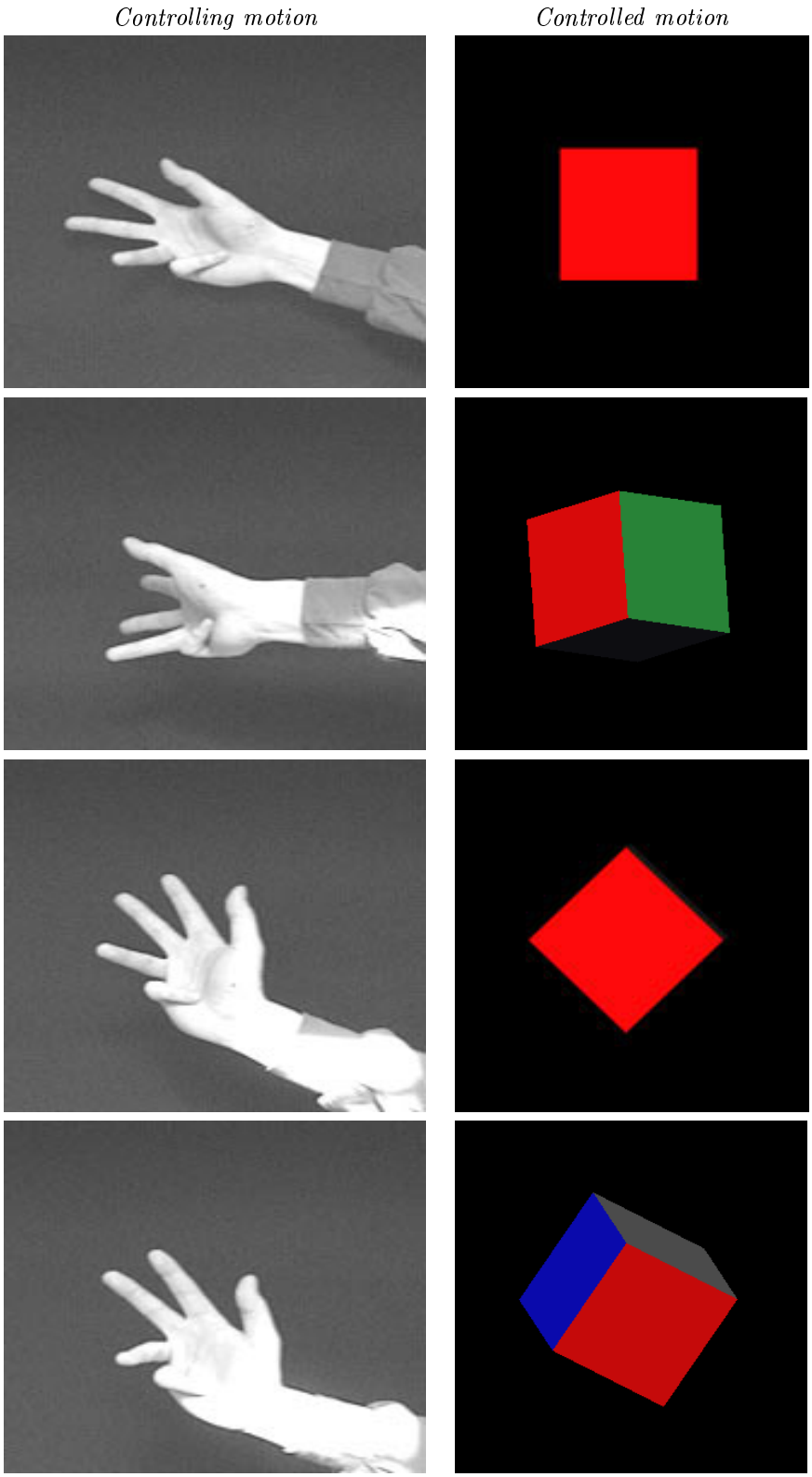

Figure 9: Illustration of the idea of a 3-D hand mouse. Here, 3-D orientation is measured from the gestures of a human hand, and is used for controlling the visualization of a cube.

regard to this application, the qualitative multi-scale feature hierarchy is a key tool for obtaining the extended feature trajectories that are needed for the subsequent structure and motion estimation.

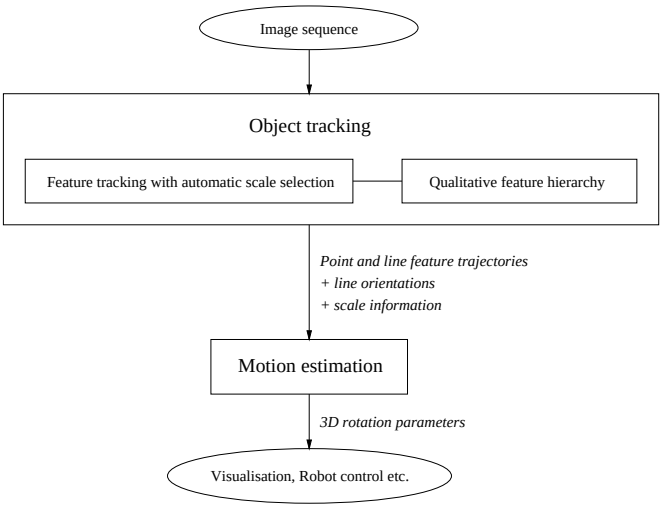

Figure 10: The components of the proposed computer vision based interface for human-computer interaction.

4.2 Sample Application II — View-Based Face Model

Figure 11 shows an example of how a qualitative feature hierarchy can support the tracking of blob features and ridge features extracted from images of a face. Again a main purpose is to recapture lost features after occlusions. Some of the detected facial features can normally be expected to change appearance over time due to e.g. blinking or mouth movements. This may cause the tracker to lose those features, but they would be recaptured as soon as the facial expression resembles the original appearance.

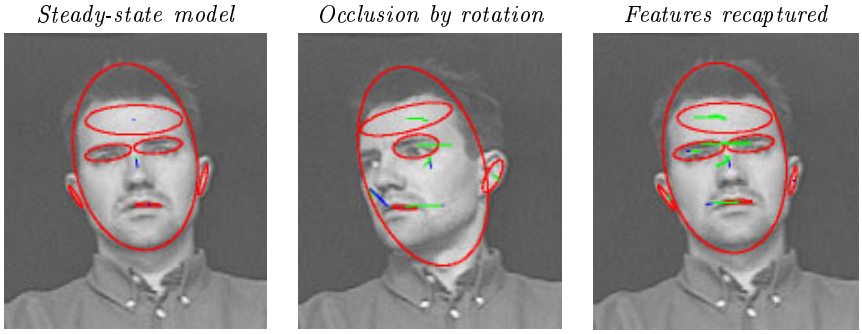

Figure 11: Results of building a qualitative feature hierarchy for a face model consisting of blob features and ridge features at multiple scales and applying this representation to the tracking of facial features over time.

5 Summary and Discussion

We have presented a view-based object representation, called the qualitative multi-scale feature hierarchy, and shown how this representation can be used for improving the performance of a feature tracker, by defining search regions in which lost features can be detected again.

Besides making explicit use of the hierarchical relations that are induced by different features in a multi-scale representation, the philosophy behind this approach is to build an internal representation that supports the processing of those image descriptors we can expect to extract from image data. This knowledge is represented in a qualitative manner, without need for constructing geometrically detailed object models.

In relation to other graph-like object representations, the discriminative power of the qualitative feature hierarchy may of course be lower than for geometrically more accurate three-dimensional object models or more detailed view-based representations involving quantitative information. Therefore, the qualitative feature hierarchies may be less suitable for object recognition, but still enough for pre-segmentation of complex scenes, or as a complement to filling in missing information given partial information from other modules (here the individual feature trackers). Notably, the application of this concept does not suffer from similar complexity problems as approaches involving explicit graph matching.

It should be pointed out that we do not claim that the proposed framework should be regarded as excluding more traditional object representations, such as three-dimensional object models or view-based representations. Rather different types of representations could be used in a complementary manner, exploiting their respective advantages. To handle the tracking of complex objects having a large number of features at the same scale level in the hierarchy, we can see several advantages in extend the proposed feature hierarchy by also defining such inter-feature relations at the same level. Such relations could be inspired by the works on labeled feature graphs by (Triesch & von der Malsburg 1996).

Concerning the determination of the qualitative relations between the features in the hierarchy, it would be interesting to explore a framework for learning the relations from training examples. In such a framework and under the assumption that our scheme for feature tracking can register features on the object that are stable over time and thereby suitable for the proposed object representation, we could consider tracking a large number of different features detected at different scales on the moving object and build up the representation either over time or a posteriori. The proposed representation is view-dependent and the result from such training sequences might indicate if more than one representation would be necessary to cover the view directions present in the sequences.

From the discussion it is evident that the proposed multi-scale feature hierarchy gives rise to a multitude of open research issues and we strongly believe that the idea should be explored further for tracking and recognition purposes.

Moreover, regarding our application to the 3-D hand mouse, it is worth pointing out that the qualitative feature hierarchy is used as a major tool in a system for computing three-dimensional structure and motion, thus at the end deriving a

quantitative three-dimensional object model from image data.

The main advantages of the proposed approach are that it is very simple to implement in practice, and that it allows us to handle semi-rigid objects, occlusions, as well as variations in view direction and illumination conditions. Specifically, with respect to the topic of scale-space theory, we have demonstrated how an integrated computer vision application with non-trivial functionally can be constructed essentially just from the following components: (i) basic scale-space operations (see section 3.1), (ii) a straightforward graph representation, and (iii) a generic framework for multi-view geometry (described elsewhere).

References

Black, M. J. & Jepson (1998*a*), A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions, *in* H. Burkhardt & B. Neumann, eds, 'Proc. 5th European Conference on Computer Vision', Vol. 1406 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Freiburg, Germany, pp. 909–924.

Black, M. J. & Jepson, A. (1998*b*), 'Eigentracking: Robust matching and tracking of articulated objects using a view-based representation', *Int. J. of Computer Vision* **26**(1), 63–84.

Bretzner, L. & Lindeberg, T. (1998*a*), 'Feature tracking with automatic selection of spatial scales', *Computer Vision and Image Understanding* **71**(3), 385–392.

Bretzner, L. & Lindeberg, T. (1998*b*), Use your hand as a 3-D mouse or relative orientation from extended sequences of sparse point and line correspondances using the affine trifocal tensor, *in* H. Burkhardt & B. Neumann, eds, 'Proc. 5th European Conference on Computer Vision', Vol. 1406 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Freiburg, Germany, pp. 141–157.

Burbeck, C. A. & Pizer, S. M. (1995), 'Object representation by cores: Identifying and representing primitive spatial regions', *Vision Research* **35**(13), 1917–1930.

Cipolla, R. & Hollinghurst, N. J. (1996), 'Human-robot interface by pointing with uncalibrated stereo vision', *Image and Vision Computing* **14**, 171–178.

Cipolla, R., Okamoto, Y. & Kuno, Y. (1993), Robust structure from motion using motion parallax, *in* 'Proc. 4th Int. Conf. on Computer Vision', Berlin, Germany, pp. 374–382.

Crowley, J. L. & Parker, A. C. (1984), 'A representation for shape based on peaks and ridges in the Difference of Low-Pass Transform', *IEEE Trans. Pattern Analysis and Machine Intell.* **6**(2), 156–170.

Crowley, J. L. & Sanderson, A. C. (1987), 'Multiple resolution representation and probabilistic matching of 2-D gray-scale shape', *IEEE Trans. Pattern Analysis and Machine Intell.* **9**(1), 113–121.

Dickinson, S. J., Pentland, A. P. & Rosenfeld, A. (1992), '3-D shape recovery using distributed aspect matching', *IEEE Trans. Pattern Analysis and Machine Intell.* **14**(2), 174–198.

Fowler, M. & Scott, K. (1997), *UML distilled*, Addison-Wesley.

Freeman, W. T. & Weissman, C. D. (1995), Television control by hand gestures, *in* 'Proc. Int. Conf. on Face and Gesture Recognition', Zurich, Switzerland.

Gauch, J. M. & Pizer, S. M. (1993), 'Multiresolution analysis of ridges and valleys in grey-scale images', *IEEE Trans. Pattern Analysis and Machine Intell.* **15**(6), 635–646.

Griffin, L. D., Colchester, A. C. F. & Robinson, G. P. (1992), 'Scale and segmentation of images using maximum gradient paths', *Image and Vision Computing* **10**(6), 389–402.

Heap, T. & Hogg, D. (1996), Towards 3D hand tracking using a deformable model, *in* 'Int. Conf. on Automatic Face and Gesture Recognition', Killington, Vermont, pp. 140–145.

Isard, M. & Blake, A. (1998), A mixed-state condensation tracker with automatic model switching, *in* 'Proc. 6th International Conference on Computer Vision', Bombay, India, pp. 107–112.

Koenderink, J. J. (1984), 'The structure of images', *Biological Cybernetics* **50**, 363–370.

Koller, D., Daniilidis, K. & Nagel, H. (1993), 'Model-based object tracking in monocular image sequences of road traffic scenes', *Int. J. of Computer Vision* pp. 257–281.

Kuch, J. J. & Huang, T. S. (1995), Vision based hand modelling and tracking for virtual teleconferencing and telecollaboration, *in* 'Proc. 5th International Conference on Computer Vision', Cambridge, MA, pp. 666–671.

Lifshitz, L. & Pizer, S. (1990), 'A multiresolution hierarchical approach to image segmentation based on intensity extrema', *IEEE Trans. Pattern Analysis and Machine Intell.* **12**(6), 529–541.

Lindeberg, T. (1993), 'Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention', *Int. J. of Computer Vision* **11**(3), 283–318.

Lindeberg, T. (1994), *Scale-Space Theory in Computer Vision*, The Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers, Dordrecht, Netherlands.

Lindeberg, T. (1998*a*), 'Edge detection and ridge detection with automatic scale selection', *Int. J. of Computer Vision* **30**(2), 117–154.

Lindeberg, T. (1998*b*), 'Feature detection with automatic scale selection', *Int. J. of Computer Vision* **30**(2), 77–116.

Lindeberg, T. & Bretzner, L. (1998), Förfarande och anordning för överföring av information genom rörelsedetektering, samt användning av anordningen. Patent pending.

Lowe, D. G. (1985), *Perceptual Organization and Visual Recognition*, Kluwer Academic Publishers, Boston.

Maggioni, C. & Kämmerer, B. (1998), Gesturecomputer-history, design and applications, *in* R. Cipolla & A. Pentland, eds, 'Computer vision for human-computer interaction', Cambridge University Press, Cambridge, U.K., pp. 23–52.

Mauerer, T. & von der Malsburg, C. (1996), Tracking and learning graphs and pose on images of faces, *in* 'Int. Conf. on Automatic Face and Gesture Recognition', Killington, Vermont, pp. 176–181.

Olsen, O. F. (1997), Multi-scale watershed segmentation, *in* J. Sporring, M. Nielsen, L. Florack & P. Johansen, eds, 'Gaussian Scale-Space Theory: Proc. PhD School on Scale-Space Theory', Kluwer Academic Publishers, Copenhagen, Denmark, pp. 191–200.

Pizer, S. M., Burbeck, C. A., Coggins, J. M., Fritsch, D. S. & Morse, B. S. (1994), 'Object shape before boundary shape: Scale-space medial axis', *J. of Mathematical Imaging and Vision* **4**, 303–313.

Shokoufandeh, A., Marsic, I. & Dickinson, S. J. (1998), View-based object matching, *in* 'Proc. 6th International Conference on Computer Vision', Bombay, India, pp. 588–595.

Smith, S. M. & Brady, J. M. (1995), 'Asset-2: Real-time motion segmentation and shape tracking', *IEEE Trans. Pattern Analysis and Machine Intell.* **17**(8), 814–820.

Toyama, K. & Hager, G. (1999), 'Incremental focus of attention for robust vision-based tracking', *Int. J. of Computer Vision* . to appear.

Triesch, J. & von der Malsburg, C. (1996), Robust classification of hand postures against complex background, *in* 'Int. Conf. on Automatic Face and Gesture Recognition', Killington, Vermont, pp. 170–175.

Vincken, K., Koster, A. & Viergever, M. (1997), 'Probabilistic multiscale image segmentation', *IEEE Trans. Pattern Analysis and Machine Intell.* **19**(2), 109–120.

Yasumuro, Y., Chen, Q. & Chihara, K. (1999), 'Three-dimensional modelling of the human hand with motion constraints', *Image and Vision Computing* **17**(2), 149–156.