Tracking of multi-state hand models using particle filtering and a hierarchy of multi-scale image features*

Ivan Laptev and Tony Lindeberg

Computational Vision and Active Perception Laboratory (CVAP),
Department of Numerical Analysis and Computer Science,
KTH, S-100 44 Stockholm, Sweden

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

Technical report CVAP245, ISRN KTH/NA/P–00/12–SE,
September 2000.

Abstract

This paper explores the use of hierarchical object representations in terms of multi-scale image features for simultaneous tracking and recognition of objects. Specifically, we consider an application to hand gesture analysis, where hand models are tracked over multiple postures (states). We propose a scale-invariant dissimilarity measure for comparing scale-space features. Based on it, we evaluate the likelihood of hierarchical, parameterized models containing different types of image features at multiple scales. The likelihood is constructed in such a way, that its maximization over different models and their parameters allows for both model selection and parameter estimation. These ideas are integrated with the framework of particle filtering, involving simultaneous tracking and recognition, and where a coarse-to-fine evaluation strategy improves computational efficiency. Based on the proposed approach, an application DrawBoard is developed, where the user controls a drawing device with a set of qualitative hand states and quantitative hand motions.

*The support from the Swedish National Board for Industrial and Technical Development, NUTEK, and the Swedish Research Council for Engineering Sciences, TFR, is gratefully acknowledged.

1 Introduction

Motivated by the inherent multi-scale nature of objects, it is natural to aim at object representations that explicitly capture the hierarchical relations between image structures at different scales. When defining such a representation, there are several possible approaches that may be taken. One alternative is to model appearance-based image patches corresponding to image structures at different levels of detail. Another approach is to extract image features at multiple scales, and to consider these as abstractions of the image structures.

While appearance based methods have been extensively applied during recent years, with a large number of successful applications, we will in this paper be concerned with object representations of the latter type. One motivation why we feel that such representations deserve further attention is that they allow us to selectively represent subsets of the image information. By choosing stable and descriptive image features of an object, we can construct a compact view-based representation for the purpose of tracking and recognizing objects under reasonable variations in the viewing conditions.

In this paper, we follow an approach of using graph-like and qualitative image representations in terms of multi-scale image features (Crowley and Sanderson 1987, Lindeberg 1993, Pizer, Burbeck, Coggins, Fritsch and Morse 1994, Triesch and von der Malsburg 1996, Siddiqi, Shokoufandeh, Dickinson and Zucker 1999, Shokoufandeh, Marsic and Dickinson 1999, Bretzner and Lindeberg 1999). Specifically, we will consider the problem of how to evaluate a hierarchical object model relative to image data, in a context where image features are extracted with an explicit mechanism for automatic scale selection. We will propose a dissimilarity measure for comparing model and data features and use this measure for evaluating the model likelihood. Then, within the paradigm of stochastic particle filtering (Isard and Blake 1996, Black and Jepson 1998, MacCormick and Blake 1999), we will show how this approach allows us to perform simultaneous alignment, tracking and recognition of hand models in different states. We will also show how the computational efficiency of particle filters can be improved by using multi-scale object models and evaluating them in a coarse-to-fine manner.

We will then apply the method to hand gesture analysis and develop an application called DrawBoard, which enables the user to control a drawing device by hand motions in a multi-functional way. By using different hand postures, the user is able to switch between different actions such as drawing, changing the pencil or zooming the drawing. At the same time, quantitative hand motions are used for moving the pencil, for changing its size and shape, as well as for rotating, translating and changing the size of the drawing.

The paper is organized as follows. We define a hand model in terms of multi-scale image features in section 2 and derive a measure for its evaluation in section 3. The particle filtering for hand tracking and recognition is then developed in section 4, and its application to hand gesture analysis is presented in section 5. Finally, a summary and conclusion is given in section 6.

2 Hand model and image features

Given an image of a hand, we can expect to detect a blob feature at a coarse scale corresponding to the palm, while fingers and finger tips may appear as ridge and blob features, respectively, at finer scales. Here, we follow the approach of feature detection

with automatic scale selection (Lindeberg 1998), and detect image features from local extrema over scales of normalized differential invariants.

2.1 Detection of image features

Given a signal f with scale-space representation $L(\cdot;\ t) = g(\cdot;\ t) * f(\cdot)$, where $g(\cdot;\ t)$ is the Gaussian kernel with variance t, a scale-space extremum of a normalized differential entity $\mathcal{D}_{norm}L$ is a point $(x;\ t)$ where $\mathcal{D}_{norm}L(x;\ t)$ assumes a local maximum or minimum with respect to space x and scale t. To detect multi-scale blobs, we search for points $(x;\ t)$ that are local extrema in scale-space of the normalized Laplacian operator

$$\nabla^2_{norm} L = t\, \nabla^2 L \tag{1}$$

and multi-scale ridges are detected as scale-space extrema of the following normalized measure of ridge strength

$$\mathcal{A}_{\gamma-norm} = t^{2\gamma}((L_{xx} - L_{yy})^2 + 4L^2_{xy}), \tag{2}$$

where $\gamma = 3/4$. Each feature detected at a point $(x,\ t)$ in scale-space indicates the presence of a corresponding image structure at position x having size t. To represent the spatial extent of such image structures, we evaluate a second moment matrix in the neighborhood of $(x;\ t)$

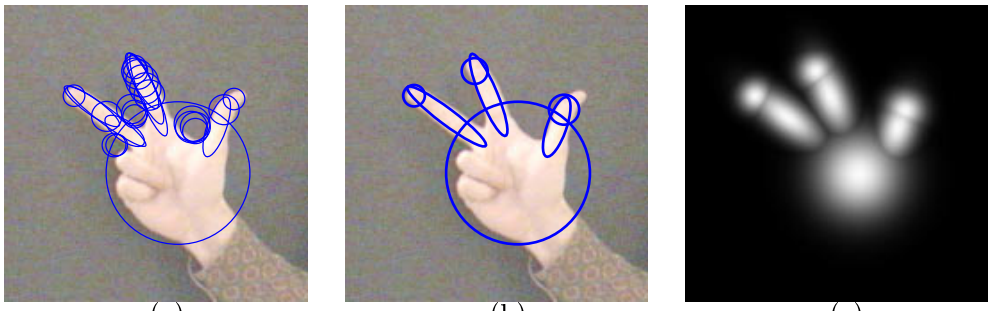
$$\nu = \int_{\eta \in \mathbb{R}^2} \begin{pmatrix} L_x^2 & L_x L_y \\ L_x L_y & L_y^2 \end{pmatrix} g(\eta; s_{int}) d\eta \tag{3}$$

at integration scale s_{int} proportional to the scale of detected features. Graphically, this image descriptor is then represented by an ellipse centered at x and with covariance matrix $\Sigma = t\nu_{norm}$, where $\nu_{norm} = \nu/\lambda_{min}$ and λ_{min} is the smallest eigenvalue of ν. Figures 1(a)-(b) show such descriptors obtained from an image of a hand.

2.2 Hierarchical and graph-like hand models

Features of objects are usually tightly coupled with each other and have stable relations in terms of their positions, scales, orientations and other attributes. We argue that such relations are important cues for recognition, since they impose additional constraints on objects and increase their discriminative power. To model such relations, it is natural to consider graph-like object representations, where the vertices in the graph correspond to features and the edges in the graph define relations between different features. Specifically, the relations between image features at different scales suggest hierarchical types of representations with coarse-scale features at the highest levels in the graph and fine-scale features at the lowest levels.

Related multi-scale representations of image structures have been proposed by several authors. Crowley and Sanderson (1987) extracted peaks from a Laplacian pyramid of an image and linked them into a tree structure with respect to their resolution. Lindeberg (1993) constructed scale-space primal sketch with an explicit encoding of blob-like structures in scale space as well as the relations between these. Triesch and von der Malsburg (1996) used elastic graphs to represent hands in different postures with local jets of Gabor filters computed at each vertex. Shokoufandeh et al. (1999) detected maxima in a multi-scale wavelet transform in a way closely related to the detection of scale-space maxima. The features were then connected into an acyclic graph according to their sizes and



(a) (b) (c)

Figure 1: Blob and ridge features for a hand: (a) circles and ellipses corresponding to the significant blob and ridge features extracted from an image of a hand; (b) selected image features corresponding to the palm, the fingers and the finger tips of a hand; (c) a mixture of Gaussian kernels associated with blob and ridge features illustrating how the selected image features capture the essential structure of a hand.

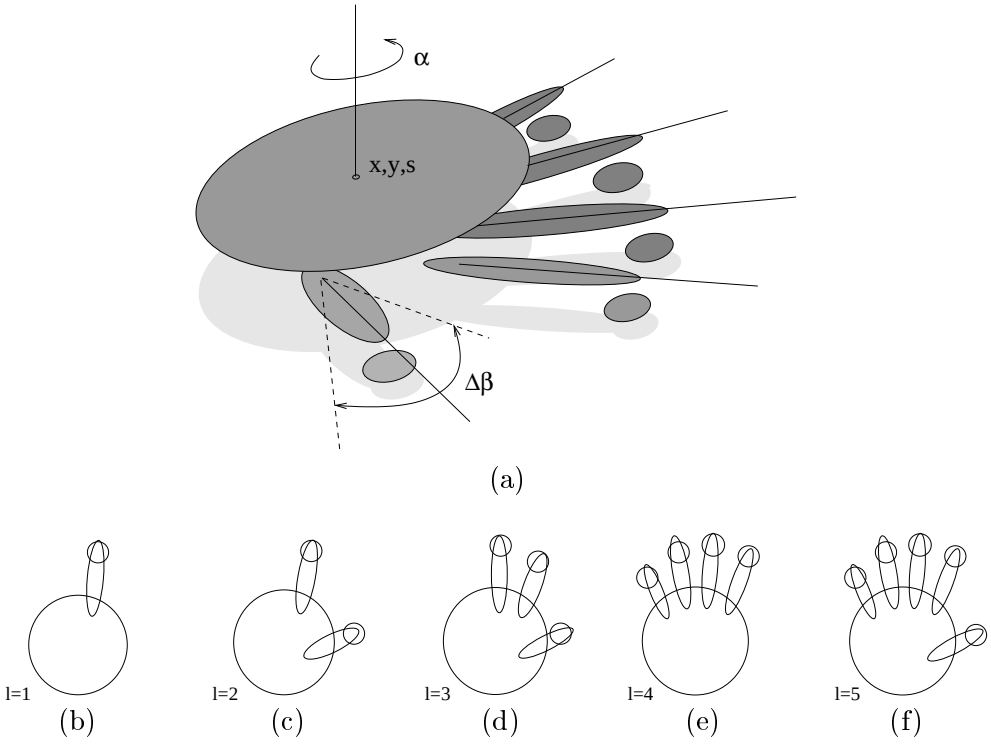


Figure 2: Model of a hand in different states: (a) hierarchical configuration of model features and their relations; (b)–(f) model states corresponding to different postures of a hand.

positions. Siddiqi et al. (1999) considered representations of binary images by shock graphs and used them for object representation and recognition. The approach by Bretzner and Lindeberg (1999) is more closely related to the one we follow here. It uses multi-scale blob and ridge features and defines explicit qualitative relations between these features across scales. Other scale-space approaches with similar aims have been developed by Lifshitz and Pizer (1990), Griffin, Colchester and Robinson (1992), Burbeck and Pizer (1995), Olsen (1997) and Vincken, Koster and Viergever (1997). Interesting works regarding hierarchical representations in biological vision has been presented by Riesenhuber and Poggio (1999).

In this paper, we will make use of quantitative relations between features to define hierarchical, probabilistic models of objects in different states. For a hand, the feature hierarchy will have three levels of detail containing a blob corresponding to a palm at the top level, ridges corresponding to fingers at the intermediate level and blobs corresponding to the finger-tips at the bottom level (see figure 2). When modeling objects in this way, a general approach is to define a prior probability density over the parameters of the all object features. In our case, we simplify this task by approximating the relative scales between all features by constant ratios and by fixing the relative positions between finger ridges and the corresponding finger-tip blobs. To capture variations in the orientation of fingers, we introduce angles β_i, $i = 1...5$ and allow for independent rotations of fingers within a fixed range $\Delta\beta$. The hand states we consider correspond to hand postures with closed and open fingers as in figures 2(b)–(f). We associate each state with an index $l = 1...5$. Finally, we describe the global position (x, y), the global scale s and the overall orientation α of the hand, which together with other parameters give a model description by (X, l) where $X = (x, y, s, \alpha, \beta_1, ..., \beta_5)$.

For our application, we manually choose model features from a set of features extracted from training images. In a more general framework, however, one could also conceive to automatically learn stable features from image data, for instance, by tracking the image features of an object over image sequences.

3 Evaluation of object model

To recognize and track hands in images, we will use a Maximum-Likelihood estimate and search for the model hypothesis (X_0, l_0) that given an image $\mathcal{I}$ maximizes the likelihood $p(\mathcal{I}|X_0, l_0)$. There are several ways of defining such a likelihood. One approach could be to relate the model features directly to local image patches. Here, we follow another idea and measure the dissimilarity between the features in the model and the features extracted from the input image.

3.1 Dissimilarity between two features

Consider image features f (here, either blobs or ridges), which are defined by positions μ and covariance matrices Σ as described in section 2.1. The dissimilarity between a pair of such features must take into account the difference in their position, size, orientation and anisotropy. Rather than treating all these attributes separately, we propose here to model the image features by two-dimensional Gaussian functions in the image domain and to compute the difference between these image representations. We use normalized

Gaussians:

$$\bar{g}(x,\mu,\Sigma) = h(\Sigma)\, g(x,\mu,\Sigma) = \frac{h(\Sigma)}{2\pi\sqrt{det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}, \tag{4}$$

where $h(\Sigma)$ is a normalizing factor which we will determine here so as to obtain a scale-invariant dissimilarity measure. The choice of Gaussian distributions is natural here, since it is the function that minimizes the entropy for a random variable given its mean and covariance. The Gaussian function at each image point can also be thought of as measuring the contribution of these points to the features. Figure 1(c) illustrates features of a hand represented by Gaussian distributions. From the similarity between this image and the original image of a hand, we note that the proposed representation actually imitates the intensity pattern of the image.

To evaluate the dissimilarity between two features f_1 and f_2, we propose to compute the square difference between their associated distributions $\bar{g}_1 = \bar{g}(x,\mu_1,\Sigma_1)$ and $\bar{g}_2 = \bar{g}(x,\mu_2,\Sigma_2)$:

$$\begin{aligned}\phi(f_1,f_2) &= \int_{\mathbb{R}^2} (\bar{g}(x,\mu_1,\Sigma_1) - \bar{g}(x,\mu_2,\Sigma_2))^2\, dx \\ &= \int_{\mathbb{R}^2} (\bar{g}_1^2 + \bar{g}_2^2 - 2\bar{g}_1\bar{g}_2)\, dx. \end{aligned} \tag{5}$$

Using the fact that the product of two Gaussian functions is another amplified Gaussian function with covariance $\hat{\Sigma} = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$ and mean $\hat{\mu} = \hat{\Sigma}(\mu_1'\Sigma_1^{-1} + \mu_2'\Sigma_2^{-1})$, i.e.

$$g(x,\mu_1,\Sigma_1)\, g(x,\mu_2,\Sigma_2) = C\, \frac{\sqrt{\det(\Sigma_1^{-1})\det(\Sigma_2^{-1})}}{2\pi\sqrt{\det(\Sigma_1^{-1}+\Sigma_2^{-1})}}\, g(x,\hat{\mu},\hat{\Sigma}), \tag{6}$$

where

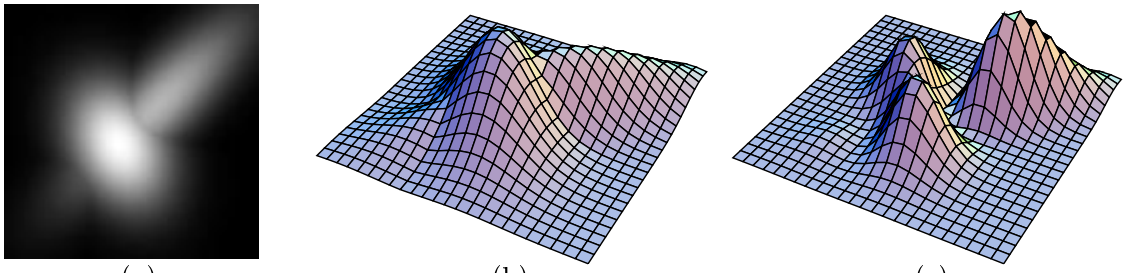
$$C = e^{-\frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 + \mu_2'\Sigma_2^{-1}\mu_2 - (\mu_1'\Sigma_1^{-1} + \mu_2'\Sigma_2^{-1})(\Sigma_1^{-1}+\Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2))},$$

the integral in (5) can be evaluated in closed form:

$$\begin{aligned}\phi(f_1,f_2) = {} & \frac{h^2(\Sigma_1)}{4\pi\sqrt{\det(\Sigma_1)}} \underbrace{\int_{\mathbb{R}^2} g(x,\mu_1,\Sigma_1/2)\, dx}_{=1} + \frac{h^2(\Sigma_2)}{4\pi\sqrt{\det(\Sigma_2)}} \underbrace{\int_{\mathbb{R}^2} g(x,\mu_2,\Sigma_2/2)\, dx}_{=1} \\ & - C\, \frac{h(\Sigma_1)h(\Sigma_2)\sqrt{\det(\Sigma_1^{-1})\det(\Sigma_2^{-1})}}{\pi\sqrt{\det(\Sigma_1^{-1}+\Sigma_2^{-1})}} \underbrace{\int_{\mathbb{R}^2} g(x,\hat{\mu},\hat{\Sigma})\, dx}_{=1}. \end{aligned} \tag{7}$$

For being useful in practice, ϕ should be invariant to the common transformations of both features. From the expression (7) it can be seen that $\phi(f_1,f_2)$ will be scale-invariant if and only if we choose $h(\Sigma) = \sqrt[4]{det(\Sigma)}$. Then we obtain

$$\phi(f_1,f_2) = \frac{1}{2\pi} - C\, \frac{\sqrt[4]{\det(\Sigma_1^{-1})\det(\Sigma_2^{-1})}}{\pi\sqrt{\det(\Sigma_1^{-1}+\Sigma_2^{-1})}}. \tag{8}$$



(a) (b) (c)

Figure 3: Two overlapping features represented by Gaussian functions in the image domain are shown (a) as a gray-value image and (b) as a three-dimensional plot. Figure (c) illustrates the square difference of these Gaussian functions, which after integration gives the dissimilarity measure between the features.

It is easy to prove that the dissimilarity measure ϕ in (8) is invariant to the common rescaling of both features, i.e. $\phi(f_1, f_2) = \phi(\tilde{f}_1, \tilde{f}_2)$, where $\tilde{f}(\mu, \Sigma) = f(\kappa\mu, \kappa^2\Sigma)$ for some scaling factor κ. It also follows that ϕ is invariant to simultaneous translation and rotation of both features.

The dissimilarity measure ϕ assumes its minimum value zero only when the features are equal, while its value increases when the features start to deviate in their positions, sizes or shapes. This idea is illustrated in figure 3. Obviously, the square difference of two Gaussian functions in figure 3(c) becomes flat for similar features, while its volume increases for features with deviating means or covariances.

3.2 Dissimilarity between model and data features

Given two sets $\mathcal{F}^m, \mathcal{F}^d$ with N^m model and N^d data features respectively, we consider the model and the data as two mixtures of Gaussian distributions in the image domain

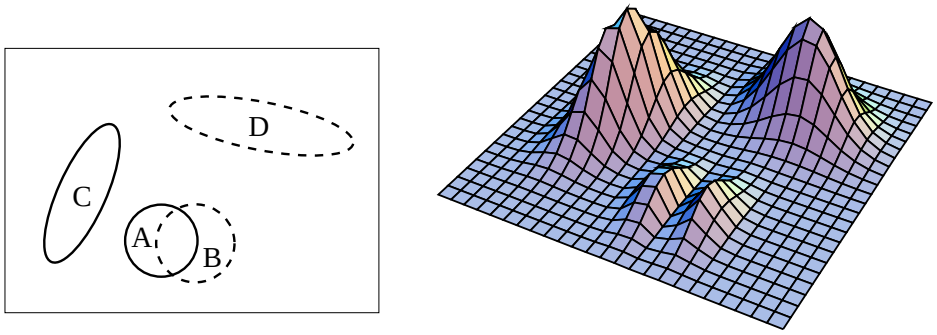
$$G^m = \sum_i^{N^m} \bar{g}(x, \mu_i^m, \Sigma_i^m), \qquad G^d = \sum_i^{N^d} \bar{g}(x, \mu_i^d, \Sigma_i^d),$$

where $\bar{g}(x, \mu_i^m, \Sigma_i^m)$ and $\bar{g}(x, \mu_i^d, \Sigma_i^d)$ are normalized Gaussian functions associated with model and data features as defined in (4). Just as for the dissimilarity between two features, we define the dissimilarity between the model and the data by integrating the square difference of their associated functions:

$$\Phi(\mathcal{F}^m, \mathcal{F}^d) = \int_{\mathbb{R}^2} (G^m - G^d)^2 \, dx. \tag{9}$$

Figure 4 illustrates the idea on two model and two data features. While the overlapping model and the data features cancel each other out, the mismatched features in both the model and the data increase the square difference $(G^m - G^d)^2$ (see figure 4(b)) and consequently increase the dissimilarity Φ. By expanding (9) we get

$$\Phi(\mathcal{F}^m, \mathcal{F}^d) = \underbrace{\sum_i^{N^m} \sum_j^{N^m} \int_{\mathbb{R}^2} \bar{g}_i^m \bar{g}_j^m \, dx}_{Q_1} + \underbrace{\sum_i^{N^d} \sum_j^{N^d} \int_{\mathbb{R}^2} \bar{g}_i^d \bar{g}_j^d \, dx}_{Q_2} - \underbrace{2 \sum_i^{N^m} \sum_j^{N^d} \int_{\mathbb{R}^2} \bar{g}_i^m \bar{g}_j^d \, dx}_{Q_3} \tag{10}$$



(a) (b)

Figure 4: Two model features (solid ellipses) and two data features (dashed ellipses) in (a) are compared by evaluating the square difference of associated Gaussian functions. While the overlapping model (A) and the data (B) features cancel each other, the mismatched features (C and D) increase the square difference in (b).

whose computation requires comparisons of all of the feature pairs. We note that overlaps between features inside $\mathcal{F}^m$ and $\mathcal{F}^d$ are rare and also not relevant for our analysis. Therefore, to save computations, we approximate the terms Q_1 and Q_2 in (10) by

$$Q_1 \approx \sum_i^{N^m} \int_{\mathbb{R}^2} (\bar{g}_i^m)^2 \, dx, \qquad Q_2 \approx \sum_i^{N^d} \int_{\mathbb{R}^2} (\bar{g}_i^d)^2 \, dx.$$

Additionally, we note that most of the model and data features will not overlap each other either and most of the products in Q_3 will be close to zero. Thus, we approximate Q_3 by

$$Q_3 \approx 2 \sum_i^{N^m} \int_{\mathbb{R}^2} \bar{g}_i^m \bar{g}_{k_i}^d \, dx,$$

where $\bar{g}_{k_i}^d$ corresponds to a data feature $f_{k_i}^d$ which is closest (according to ϕ) to the model feature f_i^m. Here, the indices $k_{1,\ldots,N^m}$ correspond to data features matched by the model, while the data features with indices $k_{N^m+1,\ldots,N^d}$ (we assume $N^d > N^m$) are considered as outliers. Taking the advantage of approximations, we finally simplify the expression (10) for Φ as

$$\begin{aligned} \Phi(\mathcal{F}^m, \mathcal{F}^d) &\approx \sum_{i=1}^{N^m} \int_{\mathbb{R}^2} ((\bar{g}_i^m)^2 - 2\bar{g}_i^m \bar{g}_{k_i}^d + (\bar{g}_{k_i}^d)^2) \, dx + \sum_{i=N^m+1}^{N^d} \int_{\mathbb{R}^2} (\bar{g}_{k_i}^d)^2 \, dx \\ &= \sum_{i=1}^{N^m} \phi(f_i^m, f_{k_i}^d) + \frac{N^d - N^m}{4\pi}, \end{aligned} \tag{11}$$

where ϕ is a dissimilarity measure of features f_i^m and $f_{k_i}^d$ according to (7). Since Φ is now expressed in terms of dissimilarities ϕ, it is clear that Φ is invariant to the common translation, rotation and re-scaling of features, just as ϕ is. The dissimilarity measure Φ is easy to compute in practice. For each model feature f_i^m, we choose a data feature $f_{k_i}^d$ that minimizes $\phi(f_i^m, f_{k_i}^d)$. Then, the sum over all $\phi(f_i^m, f_{k_i}^d)$, $i \in [1, N^m]$, together with constant terms for outliers $f_{k_i}^d$, $i \in [N^m + 1, N^d]$ give the desired measure.

By minimizing Φ according to (11), we minimize a trade-off between (i) the distance between matched model and data features (*offset criterion*) and (ii) the number of mismatched data features (*outlier criterion*). The simultaneous optimization with respect to these two criteria is important for locating an object and recognizing its state. To illustrate this property, consider an image of a hand as shown in figure 1(a) and a hand model in the different states $l = 1, 2, 3$ (see figure 2(b)-(d)). When we match such a model according to the offset criterion only, hypotheses with one or two open fingers ($l = 1, 2$) will have the same fitting error as a hypothesis with three open fingers ($l = 3$). Thus, the offset criterion alone is not sufficient for the correct selection of a hand state. To solve the problem, it is intuitive to require the best hypothesis to also explain as much of the data as possible. By adding a penalty term, measuring the number of mismatched data features (outlier criterion), we get a hypothesis that best *fits* and *explains* the data, i.e. the hypothesis with the correct state $l = 3$.

To emphasize the benefit of using multi-scale features in this context, we note that image representation in terms of such features enables for counting outliers and in this way makes it easy to estimate how well does some hypothesis explain the data. The same criterion seems to be more difficult to evaluate when comparing model hypotheses directly with intensity patterns in the image.

Currently, all the features in the model contribute equally to the dissimilarity measure Φ. By modeling the uncertainty of features and multiplying their associated Gaussian functions by the value of this uncertainty, one could decrease the influence of uncertain features and facilitate the matching to noisy data.

3.3 Likelihood

To find the best hypothesis of a hand $M(X_0, l_0)$, we must search for the minimum of Φ over X and l. For the purpose of tracking (using particle filtering as described in section 4), it is more convenient, however, to maximize a likelihood measure $p(\mathcal{I}|X, l) = p(\mathcal{F}^d|\mathcal{F}^m)$ instead. Thus, we define a likelihood function in terms of Φ as

$$p(\mathcal{F}^d|\mathcal{F}^m) = e^{-\Phi^2(\mathcal{F}^m, \mathcal{F}^d)/2\sigma^2}, \tag{12}$$

where the parameter σ^2 is estimated manually from experiments.

4 Simultaneous tracking and recognition

In this work, we are interested in tracking and recognizing a set of object models in time-dependent images. This problem can be formulated as the maximization of a posterior probability distribution over model parameters given a sequence of input images. Since exhaustive search in high-dimensional parameter spaces is usually impractical, several approaches have been developed in order to narrow down the search. Uni-modal approaches, such as Kalman filters, propagate a single hypothesis only and are known to fail in cluttered environments. The main reason of failures is due to the fact that the posterior distributions are usually multi-modal with many spurious local maxima that may lock the tracker on a wrong target. Beside this, the uni-modal approaches usually require external initialization.

Other approaches employ particle filtering in order to estimate, represent and propagate arbitrary posterior distributions over time. These approaches consider a large number

of hypotheses (particles) simultaneously while concentrating them around the maxima in distributions. Isard and Blake (1996) developed the Condensation algorithm and introduced particle filtering for visual tracking, and applied it to the problem of non-rigid contour tracking on cluttered background. Black and Jepson (1998) and Isard and Blake (1998b) used Condensation for tracking and recognizing different types of motions with application to temporal gesture recognition. MacCormick and Blake (1999) extended Condensation for tracking multiple objects. Sidenbladh, Black and Fleet (2000) and Deutscher, Blake and Reid (2000) applied particle filtering to track three-dimensional articulated person models.

Here, we will improve on particle filters in two ways. By using the multi-scale feature models and the proposed measure of their likelihood, we demonstrate how particle filters easily extend to the simultaneous tracking and recognition of objects with different spatial configurations. The second extension concerns with improvement in performance. By using hierarchical models and a coarse-to-fine strategy for their evaluation, we show how to increase the speed of the algorithm (here by a factor two) by reducing the number of particles.

4.1 Particle filtering

Particle filters aim at estimating and propagating the posterior probability distribution $p(X_t, Y_t|\mathcal{I}_t)$ over time, where X_t and Y_t are static and dynamic model parameters and $\mathcal{I}_t$ is the observation at time t. Using Bayes rule and a first-order Markov assumption, the posterior at time t can be evaluated according to

$$p(X_t, Y_t|\mathcal{I}_t) = k\, p(\mathcal{I}_t|X_t, Y_t)\, p(X_t, Y_t|\mathcal{I}_{t-1}), \tag{13}$$

where k is a normalization constant and $p(\mathcal{I}_t|X_t, Y_t)$ is the likelihood that a hypothetical model configuration X_t, Y_t gives rise to the image $\mathcal{I}_t$. The model prior $p(X_t, Y_t|\mathcal{I}_{t-1})$ conditioned on the observation in the previous time moment $\mathcal{I}_{t-1}$ is obtained from

$$p(X_t, Y_t|\mathcal{I}_{t-1}) = \int p(X_t, Y_t|X_{t-1}, Y_{t-1})\, p(X_{t-1}, Y_{t-1}|\mathcal{I}_{t-1})\, dX_{t-1}\, dY_{t-1}, \tag{14}$$

where $p(X_{t-1}, Y_{t-1}|\mathcal{I}_{t-1})$ is a posterior from a previous time moment and $p(X_t, Y_t|X_{t-1}, Y_{t-1})$ is the prediction of the model state at time t given its state at $t-1$.

Since the likelihood functions, and consequently the posterior distributions, are usually multi-modal and cannot be expressed in closed form, the approach of particle filtering is to represent distributions by a set of N discrete points in (X, Y), called particles. The particles are weighted according to their likelihoods $p(\mathcal{I}_t|X_t, Y_t)$ and the posterior for a new time moment is derived by populating the particles with high weights and predicting them according to their dynamic models $p(X_t, Y_t|X_{t-1}, Y_{t-1})$ (see (Isard and Blake 1996) for details). In the limit when $N \to \infty$, the set of particles perfectly estimates the posterior in (13). In practice, however, a sufficiently good approximation of the posterior can be obtained by taking some finite number of particles.

4.2 Hand tracking and recognition

In order to track hands described by a static model $M(X, l)$ in section 2, we use particle filtering and estimate the posterior over model parameters by a set of particles.

Since all particles correspond to hypotheses of a hand model, we compute their likelihoods according to (12) by comparing the model features with the extracted image features as described in section 3. For the dynamics of hands, we adapt a constant velocity model described by $Y_t = (\dot{x}, \dot{y}, \dot{s}, \dot{\alpha}, \dot{\beta}_1, ..., \dot{\beta}_5)$. Given the state of a hand at time $t-1$, we define its prediction $p(X_t, l_t, Y_t | X_{t-1}, l_{t-1}, Y_{t-1})$ for the time t by the following distributions

$$\begin{aligned} p(X_t|X_{t-1}, Y_{t-1}) &= \mathcal{N}(X_{t-1} + Y_{t-1}, A), \\ p(Y_t|Y_{t-1}) &= \mathcal{N}(0, B) \\ p(l_t, |l_{t-1}) &= \begin{cases} 1-\rho & l_t = l_{t-1} \\ \rho/4 & l_t \neq l_{t-1}. \end{cases} \end{aligned} \tag{15}$$

The first two distributions describe the evolution for the continuous model parameters X and Y. For simplicity reasons, we approximate all the parameters in X and Y as independent. This implies that all the off-diagonal elements of the covariance matrices A and B are zero. We adjust the variances manually by estimating them from experiments. The last equation in (15) describes the changes of a discrete model state l. According to it, the state of ρN samples (we take $\rho = 0.3$) is chosen randomly at each time moment. In this way, we enable the capture of changing hand postures in images.

When starting the tracking, we uniformly distribute all particles in (X, l, Y) and then resample and predict them and evaluate their likelihoods for every new image in a sequence. For each time moment, we estimate the best hypothesis of a hand $M(X_0, l_0)$ by first choosing its state l_0 and then computing the mean of the posterior distribution for that state. We choose $l_0 = i$ if $w_i = \max_j(w_j)$, $j = 1, ..., 5$, where w_j is the sum of weights of particles with the state j. Then, we estimate X_0 by taking a weighted mean of all particles in the state l_0. While X_0 describes continuous parameters of a hand such as its position, scale and orientation, l_0 determines the discrete state of a hand and provides a solution to the recognition of hand postures.

4.3 Hierarchical sampling

The number of particles that are used to represent distributions is directly connected to the speed and the accuracy of particle filters. A large number of particles increases the accuracy of a filter while decreasing its speed and vice versa. The number of particles can be reduced without loss of precision if restricting the temporal or spatial priors of a model. This, however, restricts the class of solvable problems and is often undesirable.

An alternative improvement can be achieved by breaking the evaluation of the likelihood function into several steps and reducing the search space after each step of computation. This idea was followed by several authors (Isard and Blake 1998a, Sullivan, Blake, Isard and MacCormick 1999, Deutscher et al. 2000). In this context, our approach bears high similarity to the work by MacCormick and Isard (2000), who developed an extension to particle filters called partitioned sampling. According to their approach, the set of particles is resampled after the evaluation of each partition (parameter subspace). This is in contrast to the standard approach where the resampling is done once at each time step for all model parameters. Partitioned sampling results in the increased concentration of particles around the maxima of the estimated distributions and therefore improves the performance.

Partitioned sampling, however, does not suggest how to construct the partitions. When using hierarchical models, it is natural to subdivide their parameters according to the

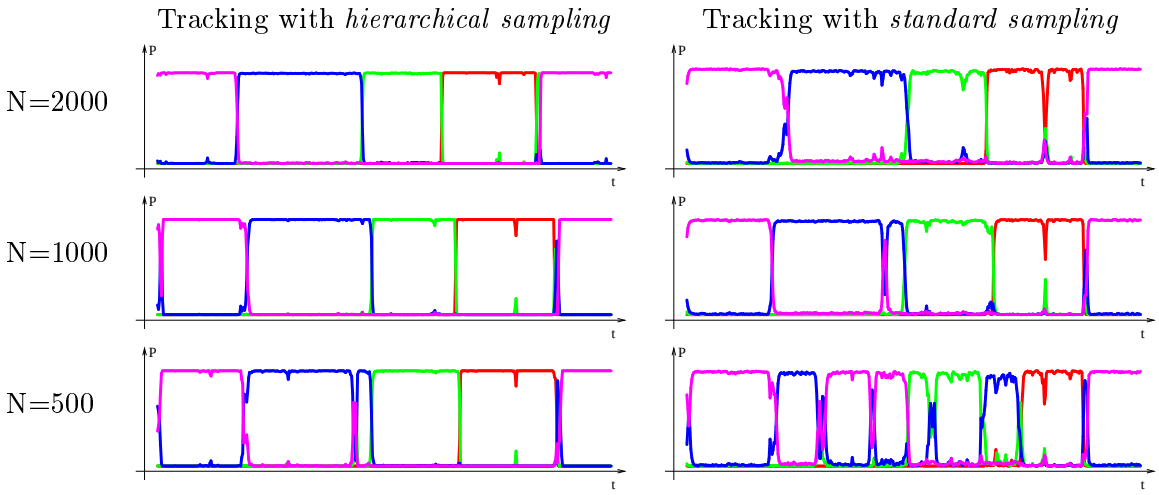


Figure 5: Curves representing probabilities of model states $l = 1, ..., 5$ while tracking a hand with changing postures. The results are shown for different numbers of used particles N and for hierarchical vs. standard sampling technique. A correct tracking and recognition is achieved with $N = 1000$ when using hierarchical sampling while the standard sampling technique required $N > 2000$.

levels of hierarchy. Different partitions shall then contain parameters that describe model structures at different scales. Using such partitions and evaluating them in the order of decreasing scale, corresponds to the commonly used coarse-to-fine search strategy. We will refer to the sampling of such partitions as *hierarchical sampling*.

Applying hierarchical sampling to hand tracking, we subdivide its parameter vector X into $X' = (x, y, s)$ and $X'' = (\alpha, \beta_1, ..., \beta_5)$. X' corresponds to the position and the scale of a hand and can be evaluated on the coarse level of the model (palm blob). X'' defines the fine structure of fingers and is evaluated on fine finger features. In our experiments, we found that subsequent resampling, prediction and evaluation of X' and X'' improves the performance of the tracker by a factor two, compared to the standard sampling method. Figure 5 illustrates this by comparing the performance of hand state recognition when tracking a hand over different postures. While the correct tracking and recognition using the hierarchical sampling required 1000 particles, the standard sampling technique required at least twice as many particles to achieve a comparable performance.

5 Application to hand gesture analysis

We are interested in tracking hands in office and home environments, in order to provide the user with a convenient interface for communicating with different kinds of computerized devices using hand gestures. The idea is to capture several degrees of freedom of a hand by a visual system and to associate them with commands in order to control devices in a multi-functional way. In particular, it may be practical to recognize discrete hand states and to associate them with different actions, while using the continuous parameters of a hand to control the actions in a quantitative way.

The problem of hand tracking and recognition has received increased attention in recent years. Some approaches consider elaborated 3D hand models and their matching to image data (Regh and Kanade 1995). Other approaches are view oriented. Cipolla,

Okamoto and Kuno (1993) used explicit color markers to simplify feature detection while tracking hand motions. Cui and Weng (1996) used appearance-based hand models for hand segmentation, tracking and sign recognition. Freeman, Tanaka, Ohta and Kyuma (1996) estimated hand states from histograms of local image orientation. Silhouettes of hands have been used for tracking by Isard and Blake (1998a), Heap and Hogg (1998) and MacCormick and Isard (2000). Maggioni and Kämmerer (1998) used gloves with special markers for human-computer interaction. Graph-like and feature-based hand models were applied by Triesch and von der Malsburg (1996) for sign recognition and by Bretzner and Lindeberg (1998) for tracking and estimating 3D rotations of a hand. While our approach relates to several these works, it is new in the respect that it simultaneously estimates continuous parameters of a hand and recognizes its discrete states during tracking.

5.1 Multi-state hand tracking

Let us first demonstrate the performance of the proposed approach when tracking hands in different states. The tracking was performed in the office environment with natural illumination variations as shown in figure 6. In the experiments, we represented and propagated the posterior distribution by $N = 1000$ particles, which were evaluated on the $N^d = 200$ strongest features extracted from input images. By using the proposed likelihood measure (12), we achieved a high discrimination of correct model states (see figure 5) despite the large number of irrelevant features in the data (see figure 1a). Figures 6(a)–(d) demonstrate the correct estimation of hand states as well as their positions, orientations and scales, when following hands with changing postures over time. The scale-invariance of the approach is illustrated in figure 6(e) where the hand is successfully tracked while undergoing considerable changes in the scale. Figure 6(f) illustrates the simultaneous estimation of rotations, scalings and translations of a hand.

The algorithm currently runs at 10Hz frame rate on a modest 550MHz Pentium III PC provided that image features are extracted off-line. A pyramid implementation of the multi-scale feature extraction should make the approach fully operational in real time in a near future.

5.2 Application to DrawBoard

To demonstrate the applicability of our approach, we developed a prototype of a drawing tool called DrawBoard, where the user is able to control a visual drawing device by a hand in a multi-functional way. A closely related approach has been previously presented by MacCormick and Isard (2000). Our approach constitutes an extension by including the interpretation of discrete hand states as well as the size of the hand in the image domain. Moreover, our approach is based on blob and ridge features, compared to the edge information used by MacCormick and Isard.

The functionality of the DrawBoard is defined as follows. When the hand is in the state $l = 1$ with one open finger (see figure 2(b)), the hand motions are used for moving the pencil over the drawing area, with the position of the pencil corresponding to the position of a finger-tip (figure 7(a)). When the user switches to a state with two open fingers (figure 2(c)), the hand movements are interpreted as drawing commands. Thus, the finger-tip now defines the position of the pencil, while the size and the orientation of the pencil is controlled by the estimated scale and the orientation of the hand (see figures 7(b)–(c) and 7(f)–(h)). When the hand is in the state $l = 3$, a rotation of the hand allows

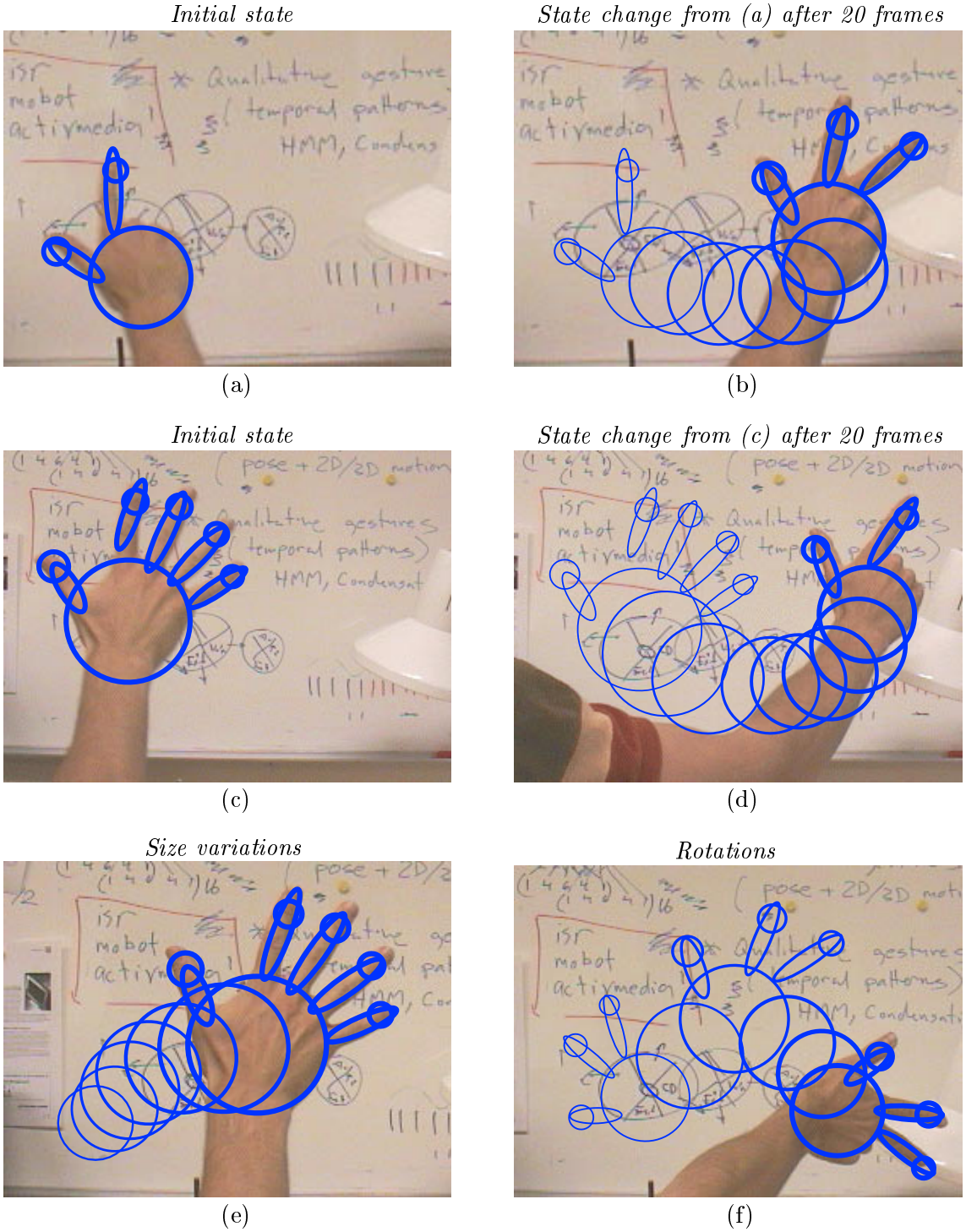


Figure 6: Result of applying the proposed framework for tracking a hand in an office environment. (a)-(b): a change in hand state from $l = 2$ to $l = 3$; (c)-(d): a change in hand state from $l = 5$ to $l = 2$; (e): size variations; (f) rotations.

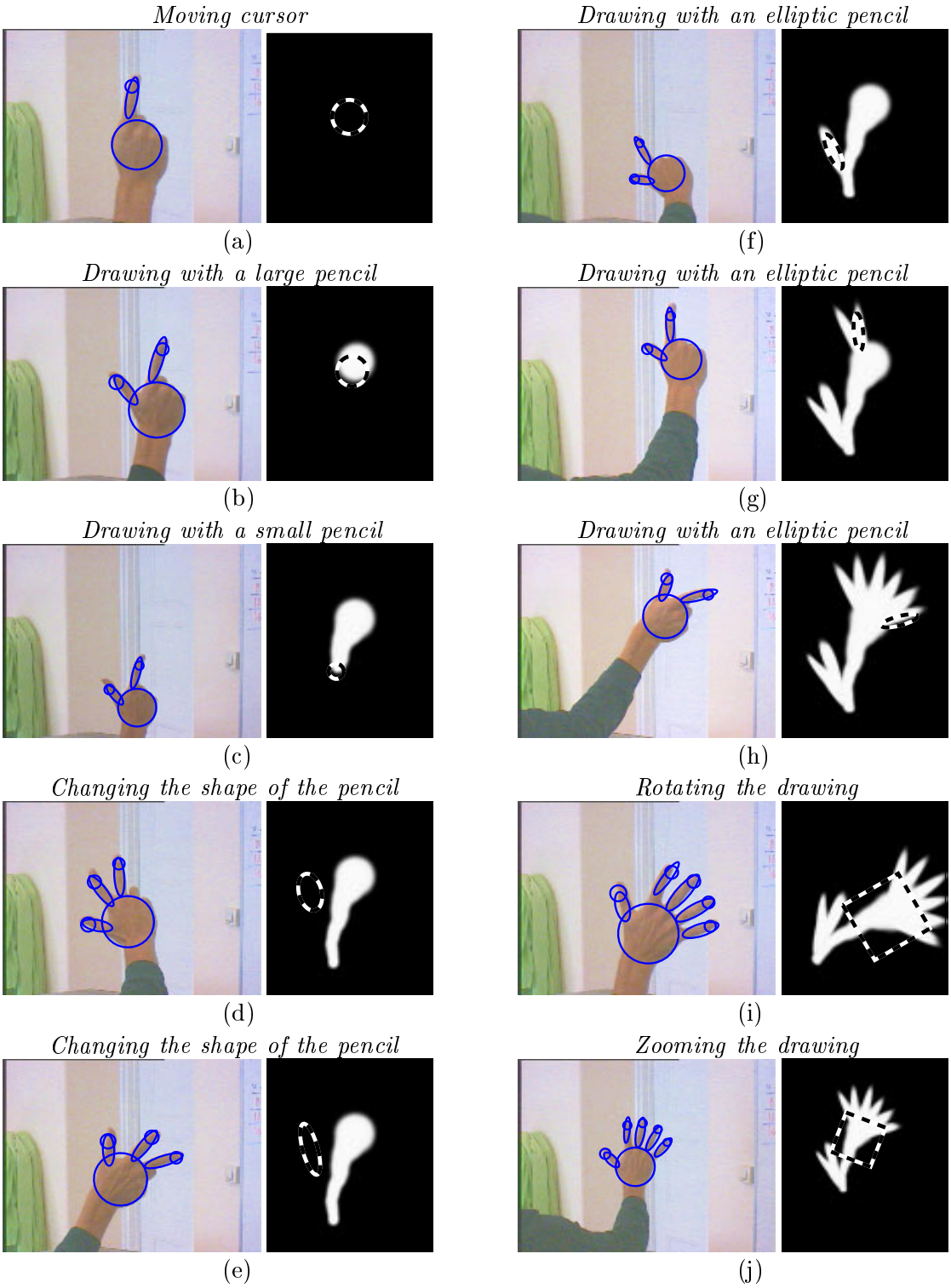


Figure 7: DrawBoard. The finger-tip in (a) is used to move the cursor. With two open fingers, the hand is used as a drawing device, whereas the size of a pencil (b)–(c) and its orientation (f)–(h) are gradually controlled by the corresponding parameters of a hand in the image. With three open fingers, the user is able to change the elliptic shape of a pencil by rotating his hand (d)–(e). The translations, scalings and rotations of the hand with five open fingers define corresponding transformations of the figure (i)–(j).

the user to gradually change the elliptic shape of the pencil by controlling the ratio of its elliptic axes (see figures 7(d)–(e)). Finally, when the hand is held in a state with five open fingers, translations, rotations and scalings of the hand are used for corresponding transformations of the figure, as illustrated on figures 7(i)–(j). The drawing produced in this way shows the simplicity and the applicability of the defined interface.[1]

While this example concerns with a drawing application, it is straightforward to redefine the interpretation of the gesture commands, and to use them for controlling other computerized equipment in an analogous fashion.

6 Summary and discussion

We have demonstrated how a view-based object representation in terms of a hierarchy of multi-scale image features can be used for tracking and recognition when integrated with the framework of particle filtering. In particular, we proposed a scale-invariant dissimilarity measure which relates such object representations with input images and enables the discrimination of different spatial configurations. The combination of this measure with multi-scale features makes our approach truly scale-invariant and allows for object tracking and recognition independently of the sizes of these objects in the image.

By applying particle filtering to hand tracking, we have shown how the quantitative characteristics of a hand as well as its qualitative states can be captured simultaneously when tracking the hand with changing postures. Specifically, the proposed technique of hierarchical sampling improves the efficiency of stochastic tracking of hierarchical multi-scale object models by reducing the number of particles.

With the application to DrawBoard we illustrated how the proposed approach can be used to construct interfaces for multi-functional control of computerized devices. In this context, the qualitative hand states are used to select between different actions, while the continuous parameters of a hand control the actions in a quantitative way.

Although this work considered hand tracking, we believe that the proposed framework should be easily extensible for tracking and recognizing of broader classes of objects. This is motivated by the compactness of object representations by a hierarchy of multi-scale features and by their efficient evaluation and discrimination using dissimilarity measures of the proposed type.

Acknowledgments

We kindly thank Lars Bretzner for many valuable suggestions concerned this work and for his help in setting up the experiments. We would also like to thank Michael Black for valuable comments and discussions based on a draft of this paper.

References

Black, M. and Jepson, A. (1998). A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions, *Fifth European Conference on Computer Vision*, Freiburg, Germany, pp. 909–924.

[1] A movie clip showing the tracking results and the operation of the DrawBoard is available from http://www.nada.kth.se/cvap/gvmdi

Bretzner, L. and Lindeberg, T. (1998). Use your hand as a 3-D mouse or relative orientation from extended sequences of sparse point and line correspondences using the affine trifocal tensor, *in* H. Burkhardt and B. Neumann (eds), *Fifth European Conference on Computer Vision*, Vol. 1406 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Freiburg, Germany, pp. 141–157.

Bretzner, L. and Lindeberg, T. (1999). Qualitative multi-scale feature hierarchies for object tracking, *in* O. F. O. M. Nielsen, P. Johansen and J. Weickert (eds), *Proc. 2nd International Conference on Scale-Space Theories in Computer Vision*, Vol. 1682, Springer Verlag, Corfu, Greece, pp. 117–128.

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

Cipolla, R., Okamoto, Y. and Kuno, Y. (1993). Robust structure from motion using motion parallax, *Fourth International Conference on Computer Vision*, Berlin, Germany, pp. 374–382.

Crowley, J. and Sanderson, A. (1987). Multiple resolution representation and probabilistic matching of 2-d gray-scale shape, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **9**(1): 113–121.

Cui, Y. and Weng, J. (1996). View-based hand segmentation and hand-sequence recognition with complex backgrounds, *13th International Conference on Pattern Recognition*, Vienna, Austria, pp. 617–621.

Deutscher, J., Blake, A. and Reid, I. (2000). Articulated body motion capture by annealed particle filtering, *CVPR'2000*, Hilton Head, SC, pp. II:126–133.

Freeman, W. T., Tanaka, K., Ohta, J. and Kyuma, K. (1996). Computer vision for computer games, *Proc. Int. Conf. on Face and Gesture Recognition*, Killington, Vermont, pp. 100–105.

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

Heap, T. and Hogg, D. (1998). Wormholes in shape space: Tracking through discontinuous changes in shape, *Sixth International Conference on Computer Vision*, Bombay, India, pp. 344–349.

Isard, M. and Blake, A. (1996). Contour tracking by stochastic propagation of conditional density, *Fourth European Conference on Computer Vision*, Cambridge, UK, pp. I:343–356.

Isard, M. and Blake, A. (1998a). Icondensation: Unifying low-level and high-level tracking in a stochastic framework, *Fifth European Conference on Computer Vision*, Freiburg, Germany, pp. 893–908.

Isard, M. and Blake, A. (1998b). A mixed-state condensation tracker with automatic model switching, *Proc. 6th Int. Conf. on Computer Vision*, Bombay, India, pp. 107–112.

Lifshitz, L. and 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, *International Journal of Computer Vision* **11**(3): 283–318.

Lindeberg, T. (1998). Feature detection with automatic scale selection, *Int. J. of Computer Vision* **30**(2): 77–116.

MacCormick, J. and Blake, A. (1999). A probabilistic exclusion principle for tracking multiple objects, *Seventh International Conference on Computer Vision*, Corfu, Greece, pp. 572–578.

MacCormick, J. and Isard, M. (2000). Partitioned sampling, articulated objects, and interface-quality hand tracking, *Sixth European Conference on Computer Vision*, Dublin, Ireland, pp. II:3–19.

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

Olsen, O. F. (1997). Multi-scale watershed segmentation, *in* J. Sporring, M. Nielsen, L. Florack and 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. and Morse, B. S. (1994). Object shape before boundary shape: Scale-space medial axis, *J. of Mathematical Imaging and Vision* **4**: 303–313.

Regh, J. M. and Kanade, T. (1995). Model-based tracking of self-occluding articulated objects, *Fifth International Conference on Computer Vision*, Cambridge, MA, pp. 612–617.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex, *Nature Neuroscience* **2**: 1019–1025.

Shokoufandeh, A., Marsic, I. and Dickinson, S. (1999). View-based object recognition using saliency maps, *Image and Vision Computing* **17**(5/6): 445–460.

Siddiqi, K., Shokoufandeh, A., Dickinson, S. and Zucker, S. (1999). Shock graphs and shape matching, *International Journal of Computer Vision* **35**(1): 13–32.

Sidenbladh, H., Black, M. and Fleet, D. (2000). Stochastic tracking of 3d human figures using 2d image motion, *Sixth European Conference on Computer Vision*, Dublin, Ireland, pp. II:702–718.

Sullivan, J., Blake, A., Isard, M. and MacCormick, J. (1999). Object localization by bayesian correlation, *Seventh International Conference on Computer Vision*, Corfu, Greece, pp. 1068–1075.

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

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