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

Abstract. This paper presents an approach for simultaneous tracking and recognition of hierarchical object representations in terms of multi-scale image features. A scale-invariant dissimilarity measure is proposed for comparing scale-space features at different positions and scales. Based on this measure, the likelihood of hierarchical, parameterized models can be evaluated in such a way that maximization of the measure over different models and their parameters allows for both model selection and parameter estimation. Then, within the framework of particle filtering, we consider the area of hand gesture analysis, and present a method for simultaneous tracking and recognition of hand models under variations in the position, orientation, size and posture of the hand. In this way, qualitative hand states and quantitative hand motions can be captured, and be used for controlling different types of computerised equipment.

1 Introduction

When representing real-world objects, an important constraint originates from the fact that different types of image features will usually be visible depending on the scale of observation. Thus, when building object models for recognition, it is natural to consider hierarchical object models that explicitly encode features at different scales as well as hierarchical relations over scales between these.

The purpose of this paper is to address the problem of how to evaluate such hierarchical object models with respect to image data. Specifically, we will be concerned with graph-like and qualitative image representations in terms of multi-scale image features (Crowley and Sanderson 1987, Lindeberg 1993, Pizer

* The support from the Swedish Research Council for Engineering Sciences, TFR, the Royal Swedish Academy of Sciences and the Knut and Alice Wallenberg Foundation is gratefully acknowledged. We also thank Lars Bretzner for many valuable suggestions concerning this work and for his help in setting up the experiments.

** Shortened version in IEEE Workshop on Scale-Space and Morphology, Vancouver, Canada, July 2001, M. Kerckhove (Ed.), Volume 2106 of Springer Verlag ¿ Lecture Notes in Computer Science, pages 63–74.

et al. 1994, Triesch and von der Malsburg 1996, Shokoufandeh et al. 1999, Bretzner and Lindeberg 1999), which are expressed within a context of feature detection with automatic scale selection. A dissimilarity measure will be proposed for comparing such model features to image data, and we will use this measure for evaluating the likelihood of object models.

Then, within the paradigm of stochastic particle filtering (Isard and Blake 1996, Black and Jepson 1998, MacCormick and Isard 2000), we will show how this approach allows us to simultaneously align, track and recognise hand models in multiple states. The approach will be applied to hand gesture analysis, and we will demonstrate how a combination of qualitative hand states and quantitative hand motions captured in this way allows us to control computerised equipment.

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 an image f with scale-space representations $L(\cdot;\ t) = g(\cdot;\ t) * f(\cdot)$, constructed by convolution with Gaussian kernels $g(\cdot;\ t)$ with variance t, a scale-space maximum 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 with respect to space x and scale t. To detect multi-scale blobs, we search for points $(x;\ t)$ that are local maxima in scale-space of the normalized squared Laplacian

$$\mathcal{B}_{\gamma-norm}L = (t\,\nabla^2 L)^2 = \sum t^2\,(\partial_{xx}L + \partial_{yy}L)^2 \tag{1}$$

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

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

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 $(x;\ t)$

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

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

(a) (b) (c)

Fig. 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 the blob and ridge features, which illustrate how the selected image features capture the essential structure of a hand.

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.

An extension of this approach to colour feature detection is presented in (Sjöbergh and Lindeberg 2001).

2.2 Hierarchical and graph-like hand models

One idea that we shall explore here is to consider relations in space and over scales between such image features as an important cue for recognition. To model such relations, we shall consider graph-like object representations, where the vertices in the graph correspond to features and the edges define relations between different features. This approach continues along the works by (Crowley and Sanderson 1987) who extracted peaks from a Laplacian pyramid of an image and linked them into a tree structure with respect to resolution, (Lindeberg 1993) who constructed a 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) who used elastic graphs to represent hands in different postures with local jets of Gabor filters computed at each vertex, (Shokoufandeh et al. 1999) who detected maxima in a multi-scale wavelet transform, as well as (Bretzner and Lindeberg 1999), who computed multi-scale blob and ridge features and defined explicit qualitative relations between these features.

Specifically, 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 contain three levels of detail; a blob corresponding to a palm at the top level, ridges corresponding to the fingers at the intermediate level and blobs corresponding to the finger-tips at the bottom level (see figure 2). While a more general approach for modelling the internal state of a hand consists of modelling the probability distribution of the parameters over all object features, we will here simplify this task by approximating the relative scales

between all features by constant ratios and by fixing the relative positions between the ridges corresponding to the fingers and the blobs corresponding to the finger-tips. Thus, we model the global position (x, y) of the hand, its overall size s and orientation α. Moreover, we have a state parameter $l = 1 \ldots 5$ describing the number of open fingers present in the hand posture (see figure 2b). In this way, a hand model can be parameterised by X, where $X = (x, y, s, \alpha, l)$.

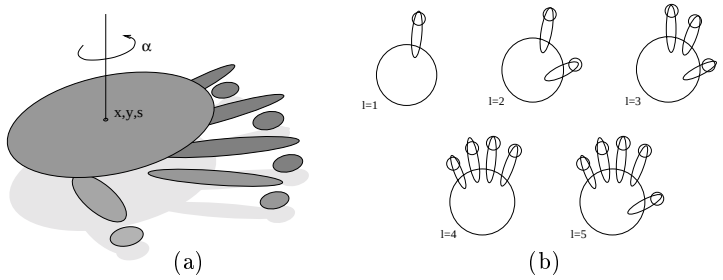


Fig. 2. Model of a hand in different states: (a) hierarchical configuration of model features and their relations; (b) model states corresponding to different hand postures.

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 that given an image $\mathcal{I}$ maximizes the likelihood $p(\mathcal{I}|X_0)$. There are several ways to define such a likelihood. One approach could be to relate the model features directly to local image patches. Here, we will measure the dissimilarity between the features in the model and the features extracted from image data.

3.1 Dissimilarity between two features

Consider an image feature F (either a blob or a ridge), defined in terms of a position μ and a covariance matrix Σ according to section 2.1. The dissimilarity between two such features must take into account the difference in their position, size, orientation and anisotropy. To measure the joint dissimilarity of these features, we propose to model each such image feature by a two-dimensional Gaussian function having the same mean and covariance as the original feature

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

and compute the integrated square difference between two such representations

$$\phi(F_1, F_2) = \int_{\mathbb{R}^2} (\bar{g}(x, \mu_1, \Sigma_1) - \bar{g}(x, \mu_2, \Sigma_2))^2 \, dx \tag{5}$$

given a normalising factor $h(\Sigma)$, which will be determined later so as to give a scale-invariant dissimilarity measure. The choice of a Gaussian function is natural here, since it is the function that minimizes the entropy of 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 this point to the image feature. Figure 1(c) illustrates features of a hand represented in this way.

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 $\hat{\mu} = \hat{\Sigma}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2)$, the integral in (5) can be evaluated in closed form:

$$\phi(F_1, F_2) = \frac{h^2(\Sigma_1)}{4\pi\sqrt{\det(\Sigma_1)}} + \frac{h^2(\Sigma_2)}{4\pi\sqrt{\det(\Sigma_2)}} - C\,\frac{h(\Sigma_1)h(\Sigma_2)\sqrt{\det(\Sigma_1^{-1})\det(\Sigma_2^{-1})}}{\pi\sqrt{\det(\Sigma_1^{-1} + \Sigma_2^{-1})}} \tag{6}$$

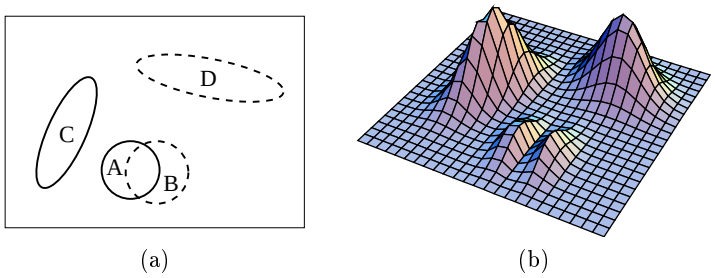
where

$$C = \exp\left(-\tfrac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 + \mu_2'\Sigma_2^{-1}\mu_2 - (\mu_1'\Sigma_1^{-1} + \mu_2'\Sigma_2^{-1})\hat{\mu})\right)$$

To be useful in practice, ϕ should be invariant to the joint translations, rotations and size variations of both features. From (6), it can be seen that $\phi(F_1, F_2)$ will be scale-invariant if and only if we choose $h(\Sigma) = \sqrt[4]{\det(\Sigma)}$. Thus, we obtain

$$\phi(F_1, F_2) = \frac{1}{2\pi} - C\,\frac{\sqrt[4]{\det(\Sigma_1^{-1})\det(\Sigma_2^{-1})}}{\pi\sqrt{\det(\Sigma_1^{-1} + \Sigma_2^{-1})}}. \tag{7}$$

It is easy to prove that the dissimilarity measure ϕ in (7) is invariant to joint rescalings 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. Moreover, ϕ is invariant to simultaneous translations and rotations of both features. As illustrated in figure 3, 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 position, size or shape.



(a) (b)

Fig. 3. 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).

3.2 Dissimilarity of 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

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

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). In analogy with the dissimilarity between two features, we define the dissimilarity between the model and the data by integrating the square difference of their associated functions:

$$\Phi(\mathcal{F}^m, \mathcal{F}^d) = \int_{\mathbb{R}^2} (G^m - G^d)^2 \, dx. \tag{8}$$

By expanding (8) we get

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

whose computation requires comparisons of all feature pairs. We can note, however, that overlap between the features within a model will be rare, as will overlaps be between features in the data. Therefore, we do the approximations

$$Q_1 \approx \sum_i^{N^m} \int_{\mathbb{R}^2} (\bar{g}_i^m)^2 \, dx, \quad Q_2 \approx \sum_i^{N^d} \int_{\mathbb{R}^2} (\bar{g}_i^d)^2 \, dx, \quad Q_3 \approx 2 \sum_i^{N^m} \int_{\mathbb{R}^2} \bar{g}_i^m \bar{g}_{k_i}^d \, dx, \tag{9}$$

where $\bar{g}_{k_i}^d$ corresponds to the data feature $F_{k_i}^d$ closest to the model feature F_i^m with regard to the dissimilarity measure ϕ. In summary, we approximate Φ by

$$\Phi(\mathcal{F}^m, \mathcal{F}^d) \approx \sum_{i=1}^{N^m} \phi(F_i^m, F_{k_i}^d) \; + \; \frac{N^d - N^m}{4\pi}, \tag{10}$$

where ϕ is the dissimilarity measure between a couple of features according to (7), F_i^m, $i = 1..N^m$ are the features of the model and $F_{k_i}^d$, $i = 1..N^m$ are the data features, where $F_{k_i}^d$ matches best with F_i^m among the other data features.

The dissimilarity measure Φ characterizes the deviation between model and data features. It is dual in the sense that it considers the distance from model features to data features (*offset criterion*) as well as the distance from data features to model features (*outlier criterion*). The simultaneous optimization with respect to these two criteria is important for locating an object and recognizing it among the others. To illustrate this, consider the matching of a hand model in states with one, two and three open fingers $l = 1, 2, 3$ (see figure 2(b)) to

an image of a hand as shown in figure 1(a). If we match according to an offset criterion only, hypotheses with one and 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, we must require the best hypothesis to also explain as much of the data as possible by minimizing the number of mismatched data features (outlier criterion). This will result in a hypothesis that best *fits* and *explains* the data, i.e. the hypothesis with the correct state $l = 3$.

3.3 Likelihood

To find the best hypothesis of a hand X_0, we will search for the minimum of the dissimilarity measure Φ in (10) over X. For the purpose of tracking (using particle filtering as will be described in section 4), it is more convenient, however, to maximize a likelihood measure $p(\mathcal{I}|X) = p(\mathcal{F}^d|\mathcal{F}^m)$ instead. Thus, we define a likelihood function as

$$p(\mathcal{F}^d|\mathcal{F}^m) = e^{-\Phi(\mathcal{F}^m,\mathcal{F}^d)/2\sigma^2}, \tag{11}$$

where the parameter $\sigma = 10^{-2}$ controls the sharpness of the likelihood function.

4 Tracking and recognition

Tracking and recognition of a set of object models in time-dependent images can be formulated as the maximization of a posterior probability distribution over model parameters given a sequence of input images. To estimate the states of object models in this respect, we will follow the approach of particle filtering to propagate object hypotheses over time, where the likelihood of each particle is computed from the proposed likelihood and dissimilarity measures (10) and (11).

To a major extent, we will follow traditional approaches for particle filtering as presented by (Isard and Blake 1996, Black and Jepson 1998, Sidenbladh et al. 2000, Deutscher et al. 2000) and others. Using the hierarchical multi-scale structure of the hand models, however, an adaptation of the layered sampling approach (Sullivan et al. 1999) will be presented, in which a coarse-to-fine search strategy is used to improve the computational efficiency, here, by a factor of two. Moreover, it will be demonstrated how the proposed dissimilarity measure makes it possible to perform simultaneous hand tracking and hand posture recognition.

4.1 Particle filtering

Particle filters aim at estimating and propagating the posterior probability distribution $p(X_t, Y_t|\tilde{\mathcal{I}}_t)$ over time, where X_t and Y_t are static and dynamic model parameters and $\tilde{\mathcal{I}}_t$ denotes the observations up to time t. Using Bayes rule, the posterior at time t can be evaluated according to

$$p(X_t, Y_t|\tilde{\mathcal{I}}_t) = k\, p(\mathcal{I}_t|X_t, Y_t)\, p(X_t, Y_t|\tilde{\mathcal{I}}_{t-1}), \tag{12}$$

where k is a normalization constant that does not depend on variables X_t,Y_t. The term $p(\mathcal{I}_t|X_t, Y_t)$ denotes the likelihood that a model configuration X_t, Y_t gives rise to the image $\mathcal{I}_t$. Using a first-order Markov assumption, the dependence on observations before time $t-1$ can be removed and the model prior $p(X_t, Y_t|\tilde{\mathcal{I}}_{t-1})$ can be evaluated using a posterior from a previous time step and the distribution for model dynamics according to

$$p(X_t, Y_t|\tilde{\mathcal{I}}_{t-1}) = \int p(X_t, Y_t|X_{t-1}, Y_{t-1})\, p(X_{t-1}, Y_{t-1}|\tilde{\mathcal{I}}_{t-1})\, dX_{t-1}\, dY_{t-1}. \quad (13)$$

Since the likelihood function is usually multi-modal and cannot be expressed in closed form, the approach of particle filtering is to approximate the posterior distribution using N particles, weighted according to their likelihoods $p(\mathcal{I}_t|X_t, Y_t)$. The posterior for a new time moment is then computed by populating the particles with high weights and predicting them according to their dynamic model $p(X_t, Y_t|X_{t-1}, Y_{t-1})$.

4.2 Hand tracking and recognition

To use particle filtering for tracking and recognition of hierarchical hand models as described in section 2, we let the state variable X denote the position (x, y), the size s, the orientation α and the posture l of the hand model, i.e., $X = (x, y, s, \alpha, l)$, while Y denotes the time derivatives of the first four variables, i.e., $Y_t = (\dot{x}, \dot{y}, \dot{s}, \dot{\alpha})$. Then, we assume that the likelihood $p(\mathcal{I}_t|X_t, Y_t)$ does not explicitly depend on Y_t, and approximate $p(\mathcal{I}_t|X_t)$ by evaluating $p(\mathcal{F}^d|\mathcal{F}^m)$ for each particle according to (11). Concerning the dynamics $p(X_t, Y_t|X_{t-1}, Y_{t-1})$ of the hand model, a constant velocity model is adopted, where deviations from the constant velocity assumption are modelled by additive Brownian motion, from which the distribution $p(X_t, Y_t|X_{t-1}, Y_{t-1})$ is computed. To capture changes in hand postures, the state parameter l is allowed to vary randomly for 30 % of the particles at each time step.

When the tracking is started, all particles are first distributed uniformly over the parameter spaces X and Y. After each time step of particle filtering, the best hypothesis of a hand is estimated, by first choosing the most likely hand posture and then computing the mean of $p(X_t, Y_t|\tilde{\mathcal{I}}_t)$ for that posture. Hand posture number i is chosen if $w_i = \max_j(w_j)$, $j = 1, \ldots, 5$, where w_j is the sum of the weights of all particles with state j. Then, the continuous parameters are estimated by computing a weighted mean of all the particles in state i.

4.3 Hierarchical layered sampling

The number of particles used for representing a distribution determines the speed and the accuracy of the particle filter. Usually, however, most of the particles represent false object hypotheses and serve as to compensate for uncertainties in the estimated distribution. To reduce the number of such particles, and thus improve the computational efficiency, one approach is to divide the evaluation

of the particles into several steps, and to eliminate unlikely particles already at the earliest stages of evaluation. This idea has been used previously in works on partitioned sampling (MacCormick and Isard 2000) and layered sampling (Sullivan et al. 1999).

The layered sampling implies that the likelihood function $p(\mathcal{I}_t|X_t)$ is decomposed as $p = p_1\, p_2 \ldots p_n$ and that false hypotheses are eliminated by re-sampling the set of particles after a likelihood $p_i(\mathcal{I}_t|X_t)$ has been evaluated at each layer $i = 1 \ldots n$. The idea is to use a coarse-to-fine evaluation strategy, where p_1 evaluates models at their coarsest scale, while p_n performs the evaluation at the finest scale.

In the context of hierarchical multi-scale feature models, the layered sampling approach can be modified such as to evaluate the likelihoods $p_i(\mathcal{I}_t|X_t)$ independently for each level in the hierarchy of features. Hence, for the hand model described in section 2, the likelihood evaluation is decomposed into three layers $p = p_1\, p_2\, p_3$, where p_1 evaluates the coarse scale blob corresponding to the palm of a hand, p_2 evaluates the ridges corresponding to the fingers, and p_3 evaluates the fine scale blobs corresponding to the finger tips.

Experimentally, we have found that the hierarchical layered sampling approach improves the computational efficiency of the tracker by a factor two, compared to the standard sampling method in particle filtering. Figure 4 illustrates a comparison between these two approaches concerning the performance of hand posture recognition step of the tracker – see (Laptev and Lindeberg 2000) for a more extensive description.

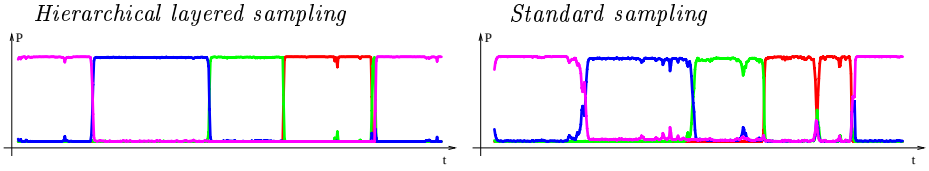


Fig. 4. Curves representing probabilities of model states $l = 1, ..., 5$ while tracking a hand with changing postures. The results are shown for the hierarchical vs. the standard sampling technique, using the same number of particles.

5 Hand gesture analysis

An application we are interested in is to track hands in office and home environments, in order to provide the user with a convenient human-machine interface for expressing commands to different types of computerized devices using hand gestures. The idea is to associate the recognised hand states with actions, while using the estimated continuous parameters of the hand model to control the actions in a quantitative way.

The problem of hand gesture analysis has received increased attention in recent years. Early work of using hand gestures for television control was presented by (Freeman & Weissman 1995) using normalised correlation. Some approaches consider elaborated 3-D hand models (Regh and Kanade 1995), while

others use colour markers to simplify feature detection (Cipolla et al. 1993). Appearance-based models for hand tracking and sign recognition were used by (Cui and Weng 1996), while (Heap and Hogg 1998, MacCormick and Isard 2000) used silhouettes of hands. Graph-like and feature-based hand models have been proposed by (Triesch and von der Malsburg 1996) for sign recognition and in (Bretzner and Lindeberg 1998) for tracking and estimating 3-D rotations of a hand.

The proposed approach is based on these works and is novel in the respect that it combines a hierarchical object model with image features at multiple scales and particle filtering for robust tracking and recognition.

5.1 Multi-state hand tracking

To investigate the proposed approach, an experiment was performed of tracking hands in different states in an office environment with natural illumination. The particle filtering was performed with $N = 1000$ particles, which were evaluated on the $N^d = 200$ strongest scale-space features extracted from each image. Figures 5(a)-(c) show a few results from this experiment. As can be seen, the combination of particle filtering with the dissimilarity measure for hierarchical object models correctly captures changes in the position, scale and orientation of the hand. Moreover, changes in hand postures are captured.

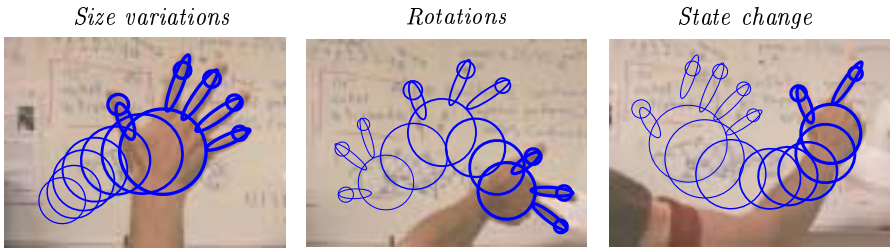


Fig. 5. Result of applying the proposed framework for tracking a hand in an office environment. (a): size variations; (b) rotations; (c): a change in hand state $l : 5 \rightarrow 2$.

As a test of the stability of the hand tracker, we developed a prototype of a drawing tool called DrawBoard, where hand motions are used for controlling a visual drawing in a multi-functional way. In this application, the cursor on the screen was controlled by the position of the hand, and depending on the state of the hand, different actions could be performed. A hand posture with two fingers implied that DrawBoard was in a drawing state, while a posture with one finger meant that the cursor moved without drawing. With three fingers present, the shape of the brush could be changed, while a hand posture with five fingers was used for translating, rotating and scaling the drawing. Figure 6 shows a few snapshots from such a drawing session.[1] As can be seen from the results, the performance of the tracker is sufficient for producing a reasonable drawing.

[1] A longer movie clip is available from http://www.nada.kth.se/cvap/gvmdi/.

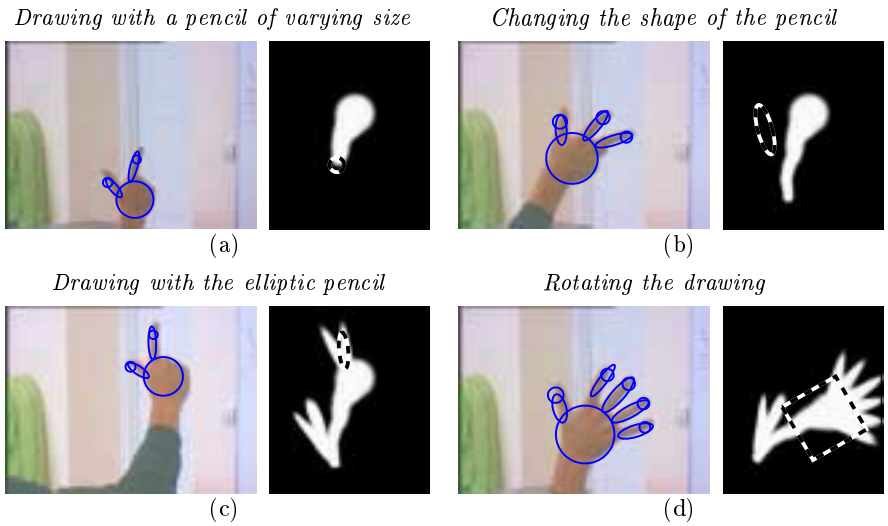


Fig. 6. DrawBoard. The hand is used as a drawing device where the position, the size and the orientation of a pencil are controlled by the corresponding parameters of a hand in the image (a),(c). In (b) the user is able to change the elliptic shape of a pencil by rotating a hand in a state with three open fingers. In (d) the drawing is scaled and rotated with a hand in a state with five open fingers.

A necessary pre-requisite for this purely intensity-based system to give satisfactory results is that there is a clear contrast in intensity between the object and the background. In on-going work, it is shown that the sensitivity to the choice of background can be reduced substantially by (i) performing colour-based feature detection, and by (ii) including a complementary prior on skin colour. In a project for computer-vision-based human-computer-interaction, this extended system is used for capturing hand gestures controlling different types of computerized equipment (Bretzner et al. 2001).

The integrated algorithm currently runs at about 10Hz frame rate on a modest dual processor PC with two 550MHz Pentium III processors. An important component in reaching real-time performance is an efficient pyramid implementation of the multi-scale feature detection step (Lindeberg and Niemenmaa 2001).

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 in combination with particle filtering, based on a scale-invariant dissimilarity measures, which relates features in the object representation to image data and enables discrimination between different spatial configurations. The combination of this measure with multi-scale features makes the approach truly scale-invariant and allows for object tracking and recognition under large size variations.

In an application to hand gesture analysis, we have shown how qualitative states and quantitative motions of a hand can be captured. In this context, the use of a hierarchical multi-scale model allows us to perform hierarchical layered sampling, which improves the computational efficiency by reducing the number of particles.

In combination with a pyramid implementation of the feature detection stage, real-time performance has been obtained, and the system has been tested in application scenarios with human-computer interaction based on hand gestures. In this context, the qualitative hand states were used for selecting between different actions, while the continuous parameters were used for controlling these actions in a quantitative way.

Although a main emphasis here has been on hand models, we believe that the proposed framework can be extended for tracking and recognizing broader classes of objects consisting of qualitatively different structures at different scales.

References

Black, M. and Jepson, A. (1998). A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions, *ECCV'98*, Freiburg, Germany, 909–924.

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, *ECCV'96*, Freiburg, Germany, 141–157.

Bretzner, L. and Lindeberg, T. (1999). Qualitative multi-scale feature hierarchies for object tracking, *Scale-Space'99*, Corfu, Greece, 117–128.

Bretzner, L., Laptev, I. and Lindeberg, T. (2001). Tracking and recognition of hand postures for visually guided control. *In preparation.*

Cipolla, R., Okamoto, Y. and Kuno, Y. (1993). Robust structure from motion using motion parallax, *ICCV'93*, Berlin, Germany, 374–382.

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

Cui, Y. and Weng, J. (1996). View-based hand segmentation and hand-sequence recognition with complex backgrounds, *ICPR'96*, Vienna, Austria, 617–621.

Deutscher, J., Blake, A. and Reid, I. (2000). Articulated body motion capture by annealed particle filtering, *CVPR'00*, Hilton Head, SC, II:126–133.

Freeman, W. T. and Weissman, C. D. (1995) , Television control by hand gestures, *Face and Gesture'95*, Zurich, Switzerland.

Heap, T. and Hogg, D. (1998). Wormholes in shape space: Tracking through discontinuous changes in shape, *ICCV'98*, Bombay, India, 344–349.

Isard, M. and Blake, A. (1996). Contour tracking by stochastic propagation of conditional density, *ECCV'96*, Cambridge, UK, I:343–356.

Laptev, I and Lindeberg, T. (2000) Tracking of multi-state hand models using particle filtering and a hierarchy of multi-scale image features, Technical report ISRN KTH/NA/P–00/12–SE. http://www.nada.kth.se/cvap/abstracts/cvap245.html

Lindeberg, T. (1993). Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention, *IJCV*, **11**: 283–318.

Lindeberg, T. (1998). Feature detection with automatic scale selection, *IJCV*, **30**(2): 77–116.

Lindeberg, T. and Niemenmaa, J. (2001). Scale selection in hybrid multi-scale representations, *in preparation*.

MacCormick, J. and Isard, M. (2000). Partitioned sampling, articulated objects, and interface-quality hand tracking, *ECCV'00*, Dublin, Ireland, II:3–19.

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, *JMIV* **4**: 303–313.

Regh, J. M. and Kanade, T. (1995). Model-based tracking of self-occluding articulated objects, *ICCV'95*, Cambridge, MA, 612–617.

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

Sidenbladh, H., Black, M. and Fleet, D. (2000). Stochastic tracking of 3-D human figures using 2-D image motion, *ECCV'00*, Dublin, Ireland, II:702–718.

Sjöbergh, J. and Lindeberg, T. (2001). *In preparation*.

Sullivan, J., Blake, A., Isard, M. and MacCormick, J. (1999). Object localization by Bayesian correlation, *ICCV'99*, Corfu, Greece, 1068–1075.

Triesch, J. and von der Malsburg, C. (1996). Robust classification of hand postures against complex background, *Face and Gesture'96*, Killington, Vermont, 170–175.