Hand Gesture Recognition using Multi-Scale Colour Features, Hierarchical Models and Particle Filtering

Lars Bretzner[1,2] Ivan Laptev[1] Tony Lindeberg[1]
[1]Computational Vision and Active Perception Laboratory (CVAP)
[2]Centre for User Oriented IT Design (CID)
Dept of Numerical Analysis and Computing Science, KTH, 100 44 Stockholm, Sweden
bretzner,laptev,tony@nada.kth.se

Abstract

This paper presents algorithms and a prototype system for hand tracking and hand posture recognition. Hand postures are represented in terms of hierarchies of multi-scale colour image features at different scales, with qualitative inter-relations in terms of scale, position and orientation. In each image, detection of multi-scale colour features is performed. Hand states are then simultaneously detected and tracked using particle filtering, with an extension of layered sampling referred to as hierarchical layered sampling. Experiments are presented showing that the performance of the system is substantially improved by performing feature detection in colour space and including a prior with respect to skin colour. These components have been integrated into a real-time prototype system, applied to a test problem of controlling consumer electronics using hand gestures. In a simplified demo scenario, this system has been successfully tested by participants at two fairs during 2001.

1 Introduction

An appealing feature of gestural interfaces is that they could make it possible for users to communicate with computerized equipment without need for external control devices, and thus e.g. replace remote controls. We have seen a number of research efforts in this area during recent years, see section 6 for an overview of works related to this one. Examples of applications of hand gesture analysis include (i) control of consumer electronics, (ii) interaction with visualization systems, (iii) control of mechanical systems and (iv) computer games.

The purpose of this work is to demonstrate how a real-time system for hand tracking and hand posture recognition can be constructed combining shape and colour cues by (i) colour feature detection in combination with qualitative hierarchical models for representing the hand and (ii) particle filtering with hierarchical sampling for simultaneous tracking and posture recognition.

Figure 1: An example of how gesture interfaces could possibly replace or complement remote controls. In this scenario, a user controls consumer electronics with hand gestures. The prototype system is described in section 5.

2 Representing the hand

The human hand is a highly deformable articulated object with many degrees of freedom and can through different postures and motions be used for expressing information for various purposes. General tracking and accurate 3D pose estimation would therefore probably require elaborate 3D hand models with time-consuming initialization and updating/tracking procedures. Our aim here is to track a number of well-defined, purposeful hand postures that the user performs in order to communicate a limited set of commands to the computer. This allows us to use a more simple, view-based shape representation, which will still be discriminatory enough to find and track a set of known hand postures in complex scenes. We therefore represent the hand by a hierarchy of stable features at different scales that captures the shape, and combine it with skin colour cues as will be described next.

Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)
0-7695-1602-5/02 $17.00 © 2002 IEEE

2.1 Multi-scale colour features

Given an image of a hand, we can expect to detect blob and ridge features at different scales, corresponding to the parts of the hand. Although the colour of the hand and the background can differ significantly, the difference in grey-level might be small and grey-level features may therefore be hard to detect on the hand. We use a recently developed approach for colour based image feature detection, based on scale-space extrema of normalized differential invariants [13]. This scheme gives more robust features than a pure grey-level based feature detection step, and consists of the following processing steps: The input RGB image is first transformed into an Iuv colour space:

$$I = \frac{R+G+B}{3} = f_1 \tag{1}$$

$$u = R - G = f_2 \tag{2}$$

$$v = G - B = f_3. \tag{3}$$

A scale-space representation is computed for each colour channel f_i by convolution with Gaussian kernels $g(\cdot;\ t)$ of different variance t, giving rise to three multi-scale colour channels $C_i(\cdot;\ t) = g(\cdot;\ t) * f_i(\cdot)$. To detect multi-scale blobs, we search for points $(x;\ t)$ that are local maxima in scale-space of the normalized squared Laplacian summed up over the colour channels at each scale

$$\mathcal{B}_{norm} C = \sum_C t^2 (\partial_{xx} C_i + \partial_{yy} C_i)^2. \tag{4}$$

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

$$\mathcal{R}_{norm} C = \sum_C t^{3/2} ((\partial_{xx} C_i - \partial_{yy} C_i)^2 + 4(\partial_{xy} C_i)^2). \tag{5}$$

To represent the spatial extent of the detected image structures, we evaluate a second moment matrix in the neighborhood of $(x;\ t)$

$$\nu = \sum_i \int_{\eta \in \mathbb{R}^2} \begin{pmatrix} (\partial_x C_i)^2 & (\partial_x C_i)(\partial_y C_i) \\ (\partial_x C_i)(\partial_y C_i) & (\partial_y C_i)^2 \end{pmatrix} g \, d\eta$$

computed at integration scale t_{int} proportional to the scale of the detected image features. The eigenvector of ν corresponding to the largest eigenvalue gives the orientation of the feature. Ellipses with covariance matrices $\Sigma = t\nu_{norm}$ represent the detected blobs and ridges in figure 2(a) and 5 for grey-level and colour images. Here $\nu_{norm} = \nu / \lambda_{min}$ and λ_{min} is the smallest eigenvalue of ν. The multi-scale feature detection is efficiently performed using an oversampled pyramid structure described in [14]. This hybrid pyramid representation allows for variable degrees of subsampling and smoothing as the scale parameter increases.

(a) (b) (c)

Figure 2: The result of computing blob features and ridge features from an image of 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.

2.2 Hierarchical hand model

The image features, together with information about their relative orientation, position and scale, are used for defining a simple but discriminative view-based object model [2]. We represent the hand by a model consisting of (i) the palm as a coarse scale blob, (ii) the five fingers as ridges at finer scales and (iii) finger tips as even finer scale blobs, see figure 3. These features are selected manually from a set of extracted features as illustrated in figure 2(a-b). We then define different states for the hand model, depending on the number of open fingers.

To model translations, rotations and scaling transformations of the hand, we define a parameter vector $X = (x, y, s, \alpha, l)$, which describes the global position (x, y), the size s, and the orientation α of the hand in the image, together with its discrete state $l = 1 \ldots 5$. The vector X uniquely identifies the hand configuration in the image and estimation of X from image sequences corresponds to simultaneous hand tracking and recognition.

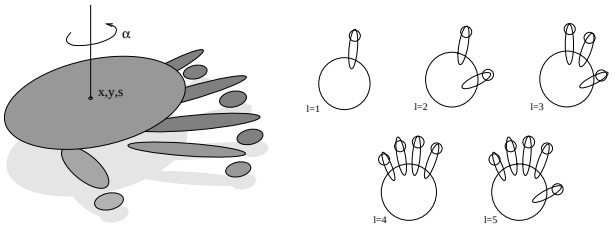


Figure 3: Feature-based hand models in different states. The circles and ellipses correspond to blob and ridge features. When aligning models to images, the features are translated, rotated and scaled according to the parameter vector X.

2.3 Probabilistic prior on skin colour

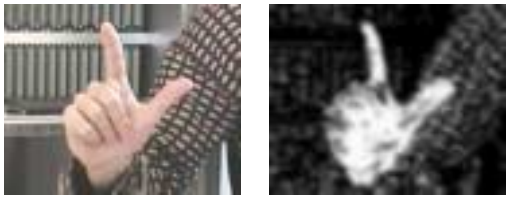
To make the hand model more discriminative in cluttered scenes, we include skin colour information in the form of a probabilistic prior, which is defined as follows:

Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)
0-7695-1602-5/02 $17.00 © 2002 IEEE

- Hands were segmented manually from the background in approximately 30 images, and two-dimensional histograms over the chromatic information (u, v) were accumulated for skin regions H_{skin}, and background H_{bg}.
- These histograms were summed up and normalized to unit mass.
- Given these training data, the probability of any measured image point with colour values (u, v) being skin colour was estimated as

$$p_{skin}(u,v) = \frac{\max(0, a\, H_{skin}(u,v) - H_{bg}(u,v))}{\sum_{u,v} \max(0, a\, H_{skin}(u,v) - H_{bg}(u,v))} \quad (6)$$

where $a = 0.1$ is a constant determining the degree of discrimination between skin colour and the colour of the background. For each hand model, this prior is evaluated at a number of image positions, given by the positions of the image features. Figure 4 shows an illustration of computing a map of this prior for an image with a hand.



(a) (b)

Figure 4: Illustration of the probabilistic colour prior. (a) original image, (b) map of the the probability of skin colour at every point.

3 Hand tracking and hand posture recognition

Tracking and recognition of a set of object models in time-dependent images can be formulated as the maximization of the a posterior probability distribution over model parameters, given a sequence of input images. To estimate the states of object models in this respect, we follow the approach of particle filtering [8, 1, 15] to propagate hypotheses of hand models over time.

3.1 Model likelihood

Particle filtering employs estimations of the prior probability and the likelihood for a set of model hypotheses. In this section we describe the likelihood function and in section 3.2 we combine it with a model prior to define a particle filter.

To evaluate the likelihood of a hand model defined in section 2.2, we compare multi-scale features of a model with the features extracted from input images. For this purpose, each feature is associated with a 2D Gaussian kernel $g(x, \mu, \Sigma)$ having the same mean μ and covariance Σ as corresponding parameters computed for image features according to section 2.1. In this way, the model and the data are represented by mixtures of Gaussians (see figure 2c) according to

$$G^m = \sum_i^{N^m} \bar{g}(x, \mu_i^m, \Sigma_i^m), \quad G^d = \sum_i^{N^d} \bar{g}(x, \mu_i^d, \Sigma_i^d), \quad (7)$$

where $\bar{g}(x, \mu, \Sigma) = \sqrt[4]{\det(\Sigma)}\, g(x, \mu, \Sigma)$. To compare the model with the data, we integrate the square difference between their associated Gaussian mixture models

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

where $\mathcal{F}^m$ and $\mathcal{F}^d$ are features of the model and the data respectively. It can be shown that this measure is invariant to simultaneous affine transformations of features. Moreover, using this measure enables for correct model selection among several models with different complexity. More details on how to compute Φ can be found in [11].

Given the dissimilarity measure Φ, the likelihood of a model hypothesis with features $\mathcal{F}^m$ on an image with features $\mathcal{F}^d$ is then estimated by

$$p(\mathcal{F}^d|\mathcal{F}^m) = e^{-\Phi(\mathcal{F}^m, \mathcal{F}^d)/2\sigma^2}, \quad (9)$$

where $\sigma = 10^{-2}$ controls the sharpness of the likelihood function. In the application to hand tracking, this entity can be multiplied by the prior $p_{skin}(u, v)$ on skin colour, described in section 2.3.

3.2 Tracking and posture recognition

Particle filters estimate and propagate 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 is evaluated according to

$$p(X_t, Y_t|\tilde{\mathcal{I}}_t) = k\, p(\mathcal{I}_t|X_t, Y_t)\, p(X_t, Y_t|\tilde{\mathcal{I}}_{t-1}), \quad (10)$$

where the prior $p(X_t, Y_t|\tilde{\mathcal{I}}_{t-1})$ and the likelihood $p(\mathcal{I}_t|X_t, Y_t)$ are approximated by the set of randomly distributed samples, i.e. hypotheses of a model and k is a normalization constant that does not depend on X_t,Y_t.

For tracking and recognition of hands, 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 approximate the likelihood $p(\mathcal{I}_t|X_t, Y_t) = p(\mathcal{I}_t|X_t)$ by evaluating the likelihood function $p(\mathcal{F}^d|\mathcal{F}^m)$ for each particle according to (9). The model prior $p(X_{t-1}, Y_{t-1}|\tilde{\mathcal{I}}_{t-1})$ restricts the dynamics of the hand and adopts a constant velocity model,

Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)
0-7695-1602-5/02 $17.00 © 2002 IEEE

where deviations from the constant velocity assumption are modeled by additive Brownian motion. 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 hand hypothesis is estimated, by first choosing the most likely hand posture and then computing the mean of $p(X_t, l_t, Y_t|\tilde{\mathcal{I}}_t)$ 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 and the weight of each particle is proportional to its likelihood. Then, the continuous parameters are estimated by computing a weighted mean of all the particles in state i. To improve the computational efficiency, the number of particles corresponding to false hypotheses are reduced using hierarchical layered sampling. The idea is related to previous works on partitioned sampling [15] and layered sampling [19]. 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. For our hand model, 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. Experiments show 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.

4 Experimental evaluation of the influence of shape and colour cues

4.1 Grey-level and colour features

A pre-requisite for a pure grey-level based feature detection system to work is that there is sufficient contrast in grey-level information between the object and the background. The first image in the first row of figure 5 shows a snapshot from a sequence with high grey-level contrast, where the hand position and pose is correctly determined using grey-level features. The grey-level features are obtained by applying the blob and ridge operators (4)–(5) to only the grey-level colour channel I in (1).

The second and third image in figure 5 show the importance of using features detected in colour space when the grey-level contrast between the object and background is low. The second image shows the detected grey-level features and how the lack of such features on the hand makes the system fail to detect the correct hand pose. The third image shows how the correct hand pose is detected using colour features. The likelihood of this situation to occur increases when the hand moves in front of a varying background.

4.2 Adding a prior on skin colour

As the number of detected features in the scene increases, so does the likelihood of hand matches not corresponding to the correct position, scale, orientation and state. In scenes with an abundance of features, the performance of the hand tracker is improved substantially by multiplying the likelihood of a model feature in (10) with this skin colour prior $p_{skin}(u, v)$. The second and third row of figure 5 shows a few snapshots from a sequence, where the hand moves in front of a cluttered background. The second row shows results without using the skin colour prior, and the third row shows corresponding results when the skin colour prior has been added. (These results were computed fully automatically; including automatic initialization of the hand model.)

Table 1 shows the results of a quantitative comparison. In a sequence of 450 frames where a moving hand changed its state four times, the result of automatic hand tracking was compared with a manually determined ground truth. While the position of the hand is correctly determined in most frames without using colour prior, the pose is often misclassified. After adding the prior on skin colour, we see a substantial improvement in both position and pose.

	no colour prior	colour prior
correct position	83%	99.5%
correct pos. and pose	45%	86.5%

Table 1: Results of a quantitative evaluation of the performance of the hand tracker in a sequence with 450 frames, with and without a prior on skin colour.

The errors in the pose estimate that remain occur spuriously, and in the prototype system described next, they are reduced by temporal filtering, at the cost of slower dynamics when capturing state changes.

5 Prototype system

The algorithms described above have been integrated into a prototype system for controlling consumer electronics with hand gestures. Figure 6 gives an overview of the system components. To increase time performance, initial detection of skin coloured regions of interest is performed, based on a wide definition of skin colour. Within these regions of interest, image features are detected using a hybrid multi-scale representation as described in section 2.1, and these image features are used as input for the particle filtering scheme outlined in section 3, with complementary use of skin colour information as described in section 2.3. On our current hardware, a dual Pentium III Xeon 550 MHz PC, this system runs at about 10 frames/s.

Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)
0-7695-1602-5/02 $17.00 © 2002 IEEE

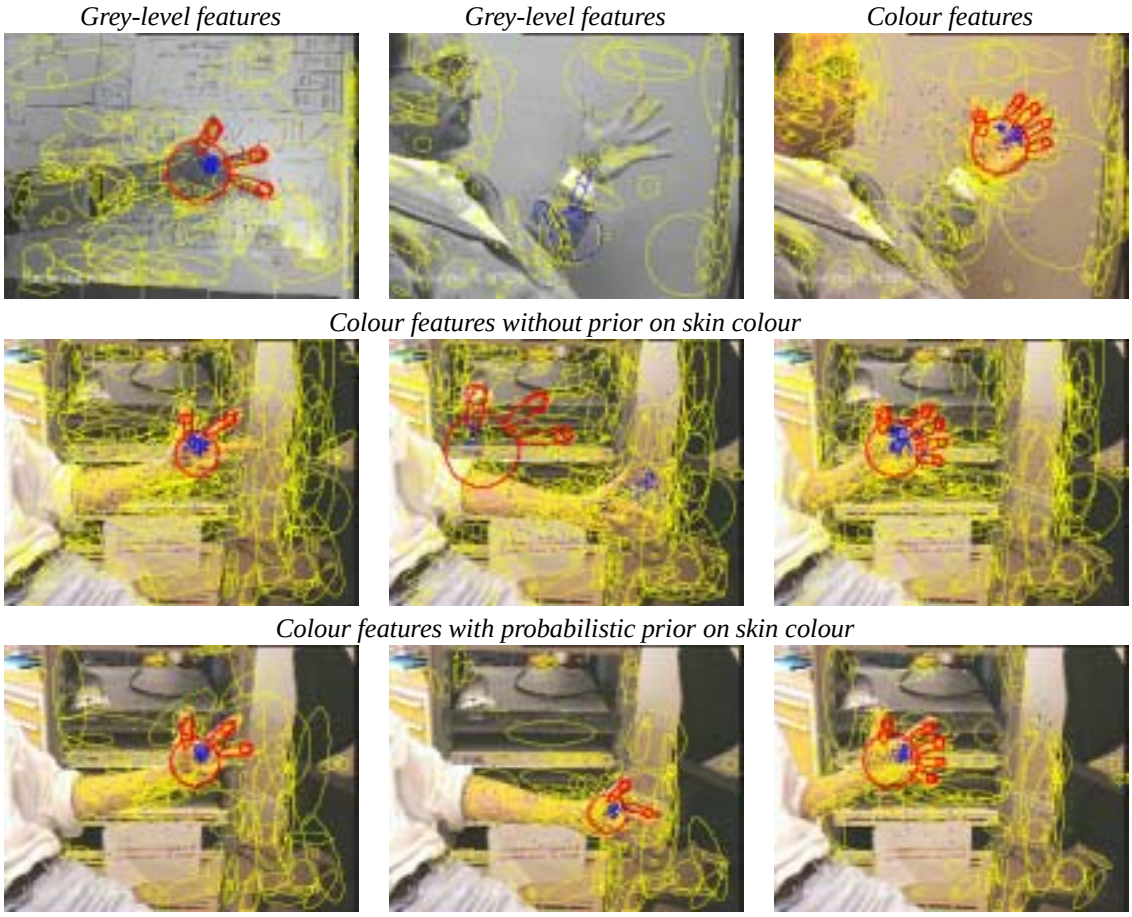


Figure 5: Illustration of the effect of combining shape and colour cues. (First row) (Left) Grey-level features are sufficient for detecting the correct hand pose when there is a clear grey-level contrast between the background and the object. (Middle, Right) When the grey-level contrast is poor, shape cues in colour space are necessary . (Middle row) With no prior on skin colour in cluttered scenes, the system often detects the wrong pose and sometimes also the wrong position. (Second row) When including this skin colour cue, both position and pose are correctly determined.

Figure 1 shows an illustration of a user who controls equipment using this system, where actions are associated with the different hand postures in the following way: Three open fingers toggle the TV on or off. Two open fingers change the channel of the TV to the next or previous depending on the rotation of the hand. Five open fingers toggle the lamp on or off. In a simplified demo scenario, this system has been presented at two IT fairs, where approximately 350 people used the system. These tests confirmed the expectations that the system, based on the described components, is user and scale (distance) invariant. To some extent the qualitative hierarchical model also shows view invariance for rotations out of the image plane (up to approx 20-30 degrees for the described gestures).

6 Related works

Early work on using hand gestures for television control was presented by Freeman and Weissman [6] using normalized correlation; see also [10, 16, 9, 21] for related works. Appearance-based models for hand tracking and sign recognition were used by Cui and Weng [4], while Heap and Hogg [7], MacCormick and Isard [15] tracked silhouettes of hands.

The use of a hierarchical hand model, continues along the works by Crowley and Sanderson [3] who extracted peaks from a Laplacian pyramid of an image and linked them into a tree structure with respect to resolution, Lindeberg [12] who 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 [20] who used elastic graphs to represent hands in different postures with local jets of Gabor filters computed

Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)
0-7695-1602-5/02 $17.00 © 2002 IEEE

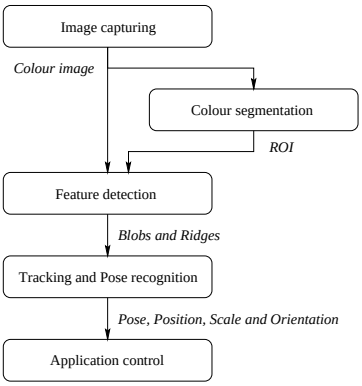


Figure 6: Overview of the main components of the prototype system for detecting and recognizing hand gestures, and using this information for controlling consumer electronics.

at each vertex, Shokoufandeh et al. [17] who detected maxima in a multi-scale wavelet transform. The use of chromaticity as a primary cue for detecting skin coloured regions was first proposed by Fleck et al. [5].

Our implementation of particle filtering largely follows the traditional approaches for condensation as presented in [8, 1, 18]. Using the hierarchical multi-scale structure of the hand models, however, we extended the layered sampling approach from Sullivan et al. [19].

7 Summary

We have presented a system for hand tracking and hand posture recognition. The main components are multi-scale colour feature hierarchies for representing hand shape, and particle filtering with hierarchical layered sampling for simultaneous tracking and recognition of hand states. In particular, we have explored the use of multi-scale colour features and probabilistic prior on skin colour. The proposed approach is novel in the respect that it combines shape and colour cues in a hierarchical object model with colour image features at multiple scales and particle filtering for robust tracking and recognition. The use of colour features gives much higher robustness to situations when there is poor grey-level contrast between the object and the background. We have also evaluated the discriminative power of including a probabilistic prior on skin colour in the particle filtering and compared the performance to the case of using colour features only. The results show that the prior on skin colour improves the discriminative power of the hand model significantly. Moreover, we have shown how these components can be integrated into a real-time prototype system for hand gesture control of computerized equipment.

References

[1] M. Black and A. Jepson. A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. In *Fifth ECCV*, pp 909–924, Freiburg, Germany, 1998.

[2] L. Bretzner and T. Lindeberg. Qualitative multi-scale feature hierarchies for object tracking. *Journal of Visual Communication and Image Representation*, 11:115–129, 2000.

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

[4] Y. Cui and J. Weng. View-based hand segmentation and hand-sequence recognition with complex backgrounds. In *13th ICPR*, pages 617–621, Vienna, Austria, 1996.

[5] M. Fleck, D. Forsyth, and C. Bregler. Finding naked people. In *Fourth ECCV*, pages II:593–602, Cambridge, UK, 1996.

[6] W. T. Freeman and C. D. Weissman. Television control by hand gestures. In *Int. Conf. on Face and Gesture Recognition*, Zurich, 1995.

[7] T. Heap and D. Hogg. Wormholes in shape space: Tracking through discontinuous changes in shape. In *Sixth International Conference on Computer Vision*, pages 344–349, Bombay, India, 1998.

[8] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *Fourth ECCV*, volume 1064 of *LNCS*, pages I:343–356, Cambridge, UK, 1996. Springer Verlag, Berlin.

[9] M. R. J. Kohler. *New contributions to vision-based human-computer-interaction in local and global environments*. PhD thesis, University of Dortmund, 1999.

[10] J. J. Kuch and T. S. Huang. Vision based hand modelling and tracking for virtual teleconferencing and telecollaboration. In *5th ICCV*, pages 666–671, Cambridge, MA, June 1995.

[11] I. Laptev and T. Lindeberg. Tracking of multi-state hand models using particle filtering and a hierarchy of multi-scale image features. In *Scale-Space'01*, vol 2106 of *LNCS*, pp 63–74. Springer, 2001.

[12] T. Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *IJCV*, 11(3):283–318, December 1993.

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

[14] T. Lindeberg and J. Niemenmaa. Scale selection in hybrid multi-scale representations. 2002. in preparation.

[15] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *Sixth European Conference on Computer Vision*, pages II:3–19, Dublin, Ireland, 2000.

[16] C. Maggioni and B. Kämmerer. Gesturecomputer-history, design and applications. In Cipolla and Pentland, *Computer vision for human-computer interaction*, pp 23–52. Cambridge University Press, 1998.

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

[18] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *Sixth European Conference on Computer Vision*, pages II:702–718, Dublin, Ireland, 2000.

[19] J. Sullivan, A. Blake, M. Isard, and J. MacCormick. Object localization by bayesian correlation. In *Seventh International Conference on Computer Vision*, pages 1068–1075, Corfu, Greece, 1999.

[20] J. Triesch and C. von der Malsburg. Robust classification of hand postures against complex background. In *Int. Conf. on Face and Gesture Recognition*, pages 170–175, Killington, Vermont, 1996.

[21] H. Watanabe et al.. Control of home appliances using face and hand sign recognition. In *Proc. 8th ICCV*, Vancouver, Canada, 2001.

Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)
0-7695-1602-5/02 $17.00 © 2002 IEEE