Active-GNG: Model Acquisition and Tracking in Cluttered Backgrounds

Anastassia Angelopoulou
School of Computer Science,
University of Westminster
Computer Vision and Imaging
Research Group
Harrow HA1 3TP, United
Kingdom
agelopa@wmin.ac.uk

Alexandra Psarrou
School of Computer Science,
University of Westminster
Computer Vision and Imaging
Research Group
Harrow HA1 3TP, United
Kingdom
psarroa@wmin.ac.uk

José García Rodríguez
School of Computer Science,
University of Alicante
Department of Computer
Technology and Computation
Apdo. 99. 03080 Alicante,
Spain
jgarcia@dtic.ua.es

Gaurav Gupta
School of Computer Science,
University of Westminster
Computer Vision and Imaging
Research Group
Harrow HA1 3TP, United
Kingdom
g.gupta1@student.wmin.ac.uk

ABSTRACT

The Self-Organising Artificial Neural Network Models, of which we have used the Growing Neural Gas (GNG) can be applied to preserve the topology of an input space. Traditionally these models neither do include local adaptation of the nodes nor colour information. In this paper, we extend GNG by adding an *active* step to the network, which we call *Active*-Growing Neural Gas (A-GNG) that has both global and local properties and can track in cluttered backgrounds. The approach is novel in that the topological relations of the model are based on a number of attributes (e.g. global and local transformations, mapping function and skin colour information) which allow us to automatically model and track $2D$ gestures. To measure the quality of the tracked correspondences we use two interlinked topology preservation measures. Experimental results have shown better performance of our proposed method over the original GNG and the Active Contour Model.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*shape, tracking*; I.5.1 [**Pattern Recognition**]: Models—*neural nets*

General Terms

Algorithms, Experimentation, Theory

VNBA'08, October 31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-313-6/08/10 ...$5.00.

Keywords

Unsupervised Learning, Tracking, Self-organising networks, Nonrigid Shapes

1. INTRODUCTION

Accurate nonrigid shape modelling and tracking is a challenging problem in machine vision with applications in human-computer interaction, motion capture, nonlinear registration, image interpretation and scene understanding. In all cases, a robust model not prone to noise that can solve for correct correspondences and can track the object of interest in a cluttered background is required. The modelling of nonrigid shapes can be performed using deformable models which can be classified as: *statistical models* and *flexible models*. From the statistical models, the most well known models are the Finite Element Models and the Statistical Shape Models. These models share in common the prior knowledge the user has about the object of interest, such as expected size, position, shape and appearance [8, 5, 12, 6, 16]. From the flexible models, the most well known models are the Hand Crafted Models, the Fourier Series Shape Models and the Active Contour Models ('Snakes'). These models share in common that no *a priori* knowledge is applied, but object-specific parameters are used as constrains for the model deformation [13, 14, 19, 17].

In all the above models, the nonrigid modelling and tracking typically involve the collection of training examples which require the segmentation and alignment of an observation sequence, which is an ill-conditioned task due to measurement noise and human variation in the observation. As a result, segmentation typically involves manual intervention and hand labelling of image sequences, a time consuming and labor intensive process that is only feasible for a limited set of gestures [12, 7]. Moreover, most of the common tracking schemes require a good representation of the posterior distribution so that low-degree parametric models can be applied to the observation [3]. This has motivated many researchers to consider nonparametric representations, including particle filters and nonparametric belief propagation [11, 18, 15].

Since we want our model to converge both globally and locally, in this paper we introduce a nonparametric approach to modelling the objects which makes it ideally suited for learning in dynamic environments. This paper presents a modification to the *Active*-GNG originally introduced in [1], which is able to (1) generalise and capture the natural variability of the objects; (2) self-organise to reflect the nonlinear correlations between inputs and outputs, thus no topological constraints; (3) track in an unsupervised manner $2D$ hand gestures in a sequence of k frames; (4) preserve topological relations based on global and local transformations, network is partially supervised, and on the distance vector between successive frames; (5) track gestures in a cluttered background by using skin colour information (6) use as few parameters as possible; (7) have a low computational complexity.

The remaining of the paper is organised as follows. Section 2 introduces the *Active* step of the GNG algorithm for the local searching and tracking of the nodes. Section 3 introduces the topology preserving measures, which are used to quantify how good are the tracked correspondences. A set of experimental results are presented in Section 4, before we conclude in Section 5.

2. *ACTIVE*-GNG

In this section, we describe A-GNG and highlight the main parts of the algorithm and its differences with the Growing Neural Gas (GNG) network introduced by Fritzke [10]. In A-GNG, nodes in the network compete for determining the set of nodes with the highest similarity to the input distribution. The highest similarity reflects which node together with its topological neighbours are nearest to the input sample point w_i which is the signal generated by the network. The n-dimensional input signals w_i are randomly generated from a finite random vector $\vec{W}$. During the learning process local error measures are gathered to determine where to insert new nodes. New nodes are inserted near the node with the highest accumulated error. At each adaptation step a connection between the winner and its topological neighbours is created as dictated by the competitive hebbian learning method. In the presence of a successive frame, the re-adaptation of the nodes is based on the current geometrical properties of the nodes, the distance vector between the original and the successive frame and the probability of the node to belong to the skin distribution.

The A-GNG network consists of:

- A set N of cluster centres known as nodes. Each node $c \in N$ has its associated reference vector $y_c \in \mathbb{R}$. The reference vectors can be regarded as positions in the input space of their corresponding nodes. Compared to the original GNG network the input space in our case can either be a discrete distribution (a finite set of all the points along the contour) or a continuous Gaussian probability density function representing skin colour. The nodes move towards the input distribution by adapting their position to the input's geometry (Figure 1). Given N reference vectors $\{\vec{y}_c\}_{c=1}^{N} \subseteq \mathbb{R}$ drawn from the random vector $\vec{W}$, we want to find a mapping $G : \mathbb{R} \longrightarrow \mathbb{R}^{\wp}$ and its inverse $F : \mathbb{R}^{\wp} \longrightarrow \mathbb{R}$ such that $\forall c = 1, ..., N$,

$$f(\vec{x}) = E_{\vec{W}|g(\vec{W})}\{\vec{W}|g(\vec{W}) = \vec{x}\}, \forall x \in N \subseteq \mathbb{R}^{\wp} \quad (1)$$

$$g(\vec{W}) = \arg \min_{\mu \in \{\vec{x}_c\}_{c=1}^{N}} \|\vec{W} - f(\mu)\| \quad (2)$$

where $\{\vec{x}_c\}_{c=1}^{N} \subseteq \mathbb{R}^{\wp}$ are the reduced reference vectors drawn from the random vector $\vec{W}$, E is the distance operator and $g(\vec{W})$ is the projection operator. Equations (1) and (2) show that while the forward mapping G is approximated as a projection operator, the reverse mapping F is nonparametric and depends on the unknown latent variable $\vec{x}$. In order to compute $f(x)$ the A-GNG algorithm evaluates (1) and (2) in an iterative manner. $\wp$ denotes the dimensionality of the latent space. In this work, current experiments include topologies of a line which is the contour of the object ($\wp = 1$) and triangular grid which is the topology preserving graph ($\wp = 2$).

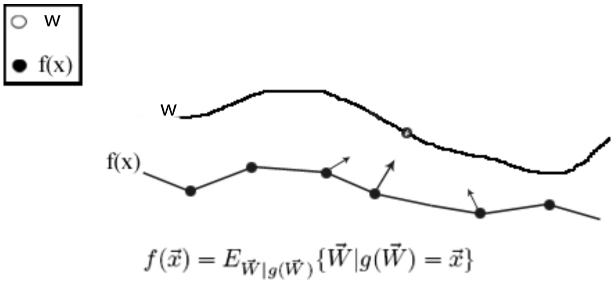
○ w
● f(x)
w
f(x)
$f(\vec{x}) = E_{\vec{W}|g(\vec{W})}\{\vec{W}|g(\vec{W}) = \vec{x}\}$

Figure 1: Every sample point w on the target space is defined as the best matching of all nodes x projecting within a topological neighbourhood of w. For example, the best matching node denoted by the largest arrow, moves towards the sample point while its topological neighbors adjust their position.

- A set A of edges (connections) between pair of nodes. These connections are not weighted and its purpose is to define the topological structure. The edges are determined using the competitive hebbian learning method. The updating rule of the algorithm is expressed as:

$$\Delta\vec{x}_{s_1} = \epsilon_x(w_i - \vec{x}_{s_1}), \Delta\vec{x}_i = \epsilon_n(w_i - \vec{x}_i) \ (\forall i \in N_{s_1}) \quad (3)$$

where ϵ_x and ϵ_n represent the constant learning rates for the winner node $\vec{x}_{s_1}$ and its topological neighbours $\vec{x}_i$, and w_i represents the randomly generated input signal from the random vector $\vec{W}$. N_{s_1} is the set of direct topological neighbours of s_1. An *edge aging scheme* is used to remove connections that are invalid due to the activation of the node during the adaptation process. Thus, the network topology is modified by removing edges not being refreshed by a time interval α_{max} and subsequently by removing the nodes connected to these edges.

The main extensions of A-GNG are summarised as follows:

1. The correspondence of the nodes is performed locally, so the model re-deforms only where differences in the

input space between successive images exist (Figure 2, 3). Therefore, the *active* step is performed locally in contrast to the global properties applied to the image by the original GNG and the 'Snake'. In the GNG example (Figure 4, 5), the topology is lost and the network collapses since the winner node $\vec{x}_{s_1}$ and its topological neighbours $\vec{x}_i$ move towards the input signal w_i. In A-GNG the topology is preserved since the updating rule $\Delta\vec{x}_{s_1}$ and $\Delta\vec{x}_i$ is performed only to the nodes where change in the input space has occurred. The rest of the network keeps its original topology structure.

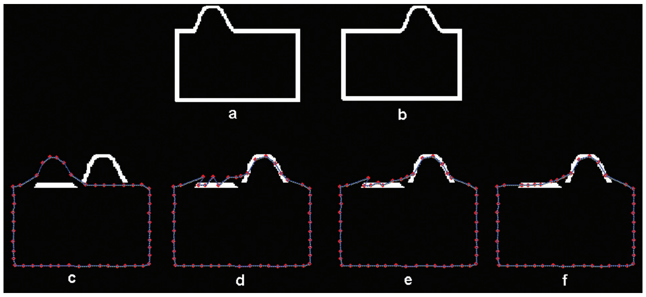

Figure 2: Example of tracking a bump model using A-GNG. a and b represent the bump model used for tracking. c - f show the local convergence of the nodes.

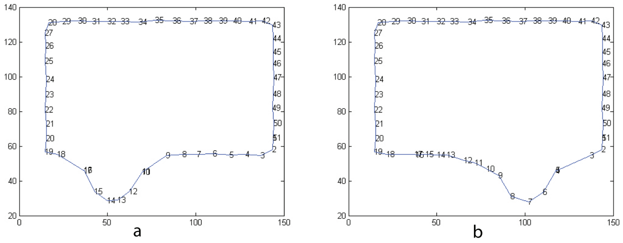

Figure 3: Image a shows the original map. In image b only nodes 10 to 17 readapt since they are the closest to the new input distribution. The rest of the map remains constant.

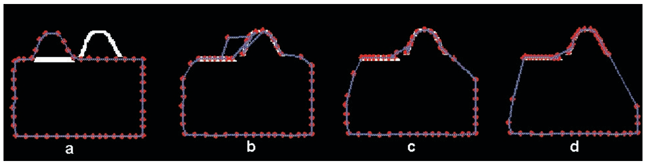

Figure 4: Tracking the same bump model using GNG. a - d show the global tracking of the nodes. The model fails to keep correspondences since all the nodes, winner node $\vec{x}_{s_1}$ and its topological neighbours $\vec{x}_i$, adjust their position according to the learning rates ϵ_x and ϵ_n. The network gradually will move all the nodes to the new input space.

2. The mean vector of the map and of any successive image is calculated and the nodes update their position based on this mean difference. By doing this the map first updates its position into the successive image and then examines a region of the image around each node to determine a better displacement of the node. Calculating the mean is very important when a rapidly moving distribution occurs. Figure 6 illustrates the ability of A-GNG to track a jumping distribution due to the calculation of the distance vector between the original and the rapidly moving distribution. In Figure 7 GNG fails to readapt to the moving distribution and the nodes of the network become inactive.

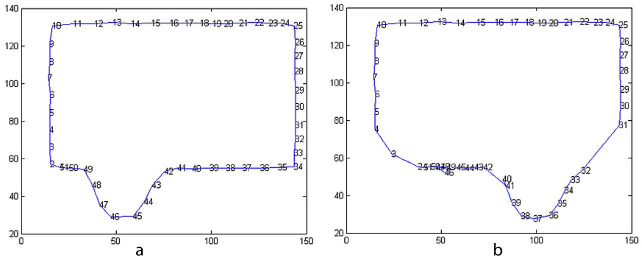

Figure 5: Image a shows the original map of the bump model. Image b shows the fold-overs that occur after a number of iterations. Fold-overs of the network occur between points 40 and 41 and between 46 and 47. Not only point correspondences are lost, but also topology relations are violated.

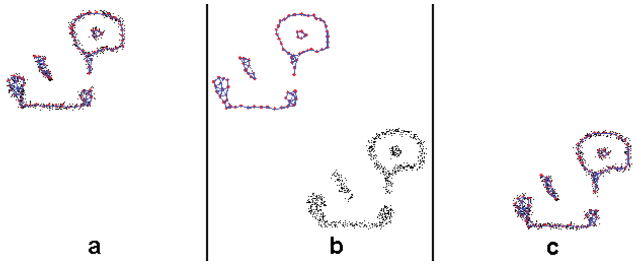

Figure 6: Three states of A-GNG tracking a nonstationary discrete jumping distribution. a - c represent the original state of the distribution and the final adaptation of the nodes to the moving distribution.

3. In order to track a hand gesture in a cluttered background we add a second dimension to the network with skin colour information taken at each node. The skin distribution derives from the standard technique of modelling a distribution using a mixture of K Gaussians [4]. If a node is represented as $(x, y, P(g(x, y)))$, where x, y are positions, $g(x, y)$ colour at x, y and $P(g(x, y))$ posterior probability of that gray level, the strength of the node can then depend on the value of the posterior probability (Figure 8).

The main steps of the A-GNG algorithm are as follows:

1. For every node calculate the probability of belonging to the skin Gaussian probability density function by using the Bayes's theorem:

$$p(k|x) = \frac{p(k)p(x|k)}{p(x)} = \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)} \qquad (4)$$

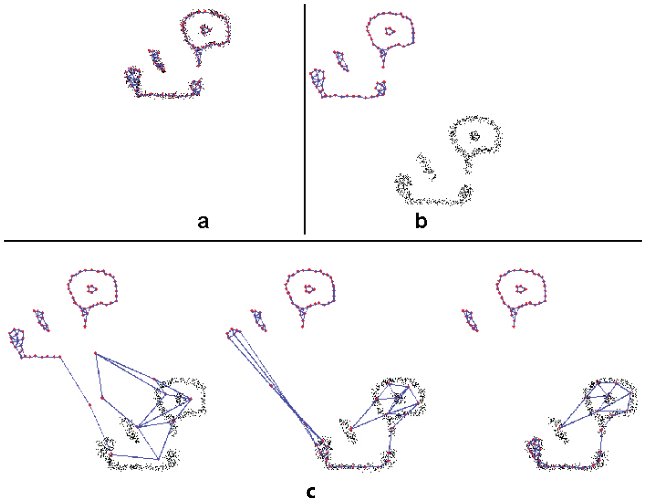

Figure 7: Three states of GNG. a and b represent the original states of the distribution and are the same with A-GNG. c represents the intermediate steps of the nodes adaptation and shows how the nodes become inactive, failing to readapt their position to the rapidly moving distribution. In this case the topology is violated and the resources of the network are wasted.

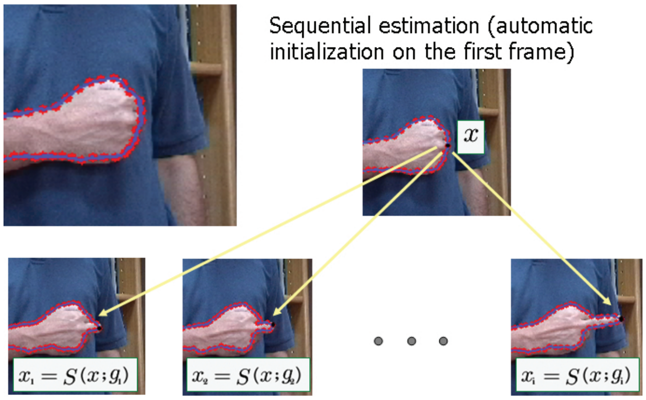
Sequential estimation (automatic initialization on the first frame)
x
$x_1 = S(x;g_1)$
$x_2 = S(x;g_2)$
• • •
$x_i = S(x;g_i)$

Figure 8: The movement of the nodes depend on the posterior probability $P(g(x,y))$. The highest the probability of a node to belong to the skin prior probability the faster the node will be adapting to the new input space.

where μ and Σ represent the mean and the covariance of the kth Gaussian. The total shape then is given as $\mathrm{S}(\vec{x}; P(g(x,y)))$ where $\vec{x}$ is a $2n(x,y)$ node vector and $P(g(x,y))$ is the posterior probability of the node at $g(x,y)$ gray level.

2. Given N number of nodes calculate the mean node $\vec{\overline{x_c}}$ of the modal image, where

$$\vec{\overline{x_c}} = \frac{1}{N}\sum_{i=1}^{N} \vec{x_i} \tag{5}$$

3. Calculate the image difference between the modal image and any other successive image. Let A and B be two sets of elements in $\mathbb{R}$ representing the modal and the successive image respectively. The Minkowski subtraction of B from A is defined as:

$$A - B = \bigcap_{b \in B} A_b \tag{6}$$

where the elements b are the pixel coordinates of the successive image.

4. Let $C = A - B$ be the new input distribution of the network.

5. Randomly generate input signals w_i to C and calculate as in step 3 the mean signal $\overline{w_i}$.

6. Calculate the distance vector of the two means and swift the nodes towards C. For each successive image we calculate its deviation from the mean, $d\vec{x_i}$ where

$$d\vec{x_i} = \vec{x_i} - \vec{\overline{x_c}} \tag{7}$$

7. Randomly generate input signals w_i to C and find the winner node x_{s_1} and its direct topological neighbours x_i.

8. Update the position of the nodes by moving them towards the current signal by the weighted factors ϵ_x, and ϵ_n same as in GNG.

9. Remove the used signal w_i from the input distribution.

10. Repeat iterations $1 - 9$ until the system converges.

3. TOPOLOGY PRESERVATION MEASURES

In order to quantify the topology preservation of the A-GNG map between tracked frames, we measure two properties of the network: its resolution and its topology preservation of the input space. To measure the resolution of the map we use the quantisation error given by:

$$E = \sum_{\forall w \in R^{\wp}} \parallel x_{s_w} - w \parallel^2 P(w) \tag{8}$$

where s_w is the nearest node to the input signal w. Quantisation error measures the accuracy of A-GNG in representing the distance, and subsequently the mapping of the different lattices, from the closest node to the input signal. Low quantisation error is desirable and indicates that neighbouring nodes of the input space must be nearest neighbours in the latent space.

To measure the topology we use the topographic product and its modification. The topographic product P introduced by Bauer and Pawelzik [2] quantifies the neighbourhood preservation of the map by computing the Euclidean distance between neighbouring nodes, in both the input and the latent space. A mapping preserves neighbourhood relations if and only if nearby points in the input space remain close in the latent space.

The neighbourhood relationship between each pair of nodes in the latent space $\wp$ and its associative reference vectors in the input space d is given by:

$$P_1(c,k) = [\prod_{l=1}^{k} \frac{d^{\wp}(c, n_l^{\wp}(c))}{d^{\wp}(c, n_l^{d}(c))}]^{1/l} \tag{9}$$

$$P_2(c,k) = [\prod_{l=1}^{k} \frac{d^d(\vec{x}_c, \vec{x}_{n_l^{\wp}(c)})}{d^d(\vec{x}_c, \vec{x}_{n_l^{d}(c)})}]^{1/l} \qquad (10)$$

where c is a node, $\vec{x}_c$ is its reference vector, n_l^d is the l-th closest neighbour to c in the input space d according to a distance d^d and $n_l^{\wp}$ is the l-th nearest node to c in the latent space $\wp$ according to a distance $d^{\wp}$. Combining (10) and (11) a measure of the topological relationship between the node c and its k closest nodes is obtained:

$$P_3(c,k) = [\prod_{l=1}^{k} \frac{d^d(\vec{x}_c, \vec{x}_{n_l^{\wp}(c)})}{d^d(\vec{x}_c, \vec{x}_{n_l^{d}(c)})} \cdot \frac{d^{\wp}(c, n_l^{\wp}(c))}{d^{\wp}(c, n_l^{d}(c))}]^{1/2k} \qquad (11)$$

To extend this measure to all the nodes of the network and all the possible neighbourhood orders, the topographic product P is defined as:

$$P = \frac{1}{N(N-1)} \sum_{c=1}^{N} \sum_{k=1}^{N-1} \log(P_3(c,k)) \qquad (12)$$

In order to use this measure to nonlinear input spaces the geodesic distance is employed as d^d in the geodesic topographic product [9]. The sign of P indicates the relationship between input and latent space. $P \approx 0$ indicates good adaptation, $P < 0$ indicates folding of $\wp$ into d, which means too low dimension of $\wp$, while $P > 0$ indicates too high dimension of $\wp$.

Figure 9 shows the graphs with the topology preservation measures, Inverse Quantisation Error (IQE), Topographic Product (TP) and the Geodesic Topographic Product (GTP) for both images. It is observed that all measures considered are suitable candidates to be used for the quantification of the topology preservation of A-GNG. The low inverse quantisation error indicates the accuracy of the map in representing its input, while the two topology measures with $P \approx 0$ indicate good adaptation. However, since the manifold of the input space is nonlinear the GTP performs better, because $P \approx 0$, compared to TP which takes negative values $P < 0$.

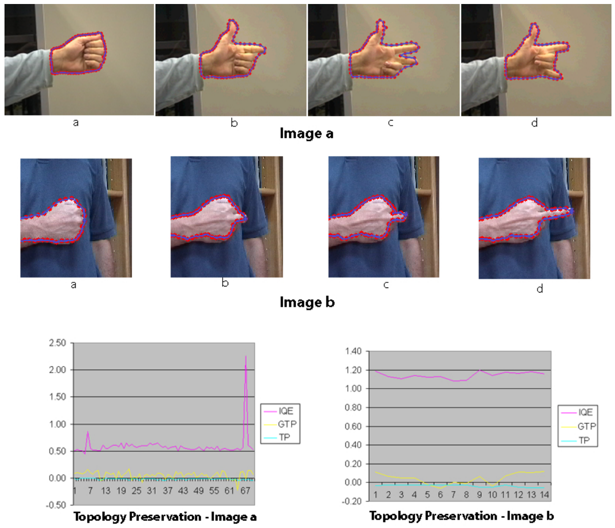

Figure 9: Topology preservation during the tracking process of A-GNG for images a and b measured with the inverse quantisation error, the topographic product and the geodesic topographic product.

4. EXPERIMENTS

We demonstrate the performance of our system on a sequence of hand gestures in different backgrounds starting from an evenly distributed gray-level background information and moving towards more cluttered backgrounds (Figure 10, 11, 12). In the following experiments we see the superiority of A-GNG against GNG which adheres only to global shape transformations.

Figure 10(a) shows the initial position of A-GNG. Images (b) and (c) show the adaptation after 1 iteration to a very subtle movement. Images (d), (f) and (h) show the updated position of the network to a more jumping distribution and how the network re-adapts again after 1 iteration. Figure 11(a) shows the initial A-GNG position. Images (b) to (i) show the tracking of the nodes to a sequence of 90 frames. Our tracker is able to track the fingers and updates the topology of the network every 5 iterations. Figure 12 indicates a $2D$ topology tracking example to a sequence of 90 frames. Images $a - h$ show the adaptation of the network after 5 iterations to a cluttered background using the skin colour distribution as the input space of the A-GNG network.

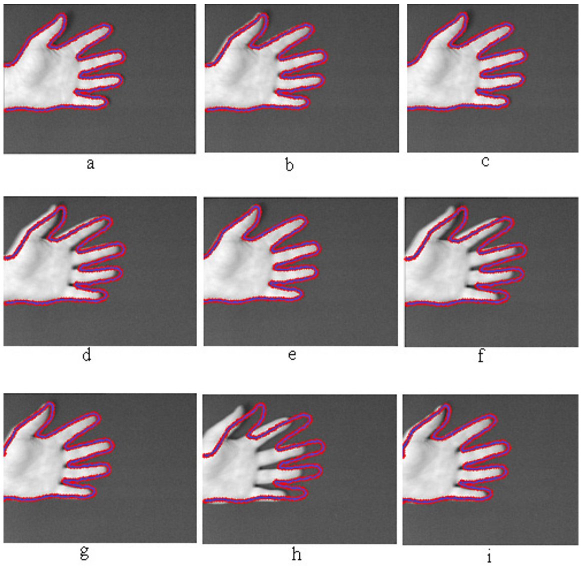

Figure 10: Tracking a hand. The images correspond from left to right and from top to bottom to every 5th frame of a 45 frame sequence.

5. CONCLUSIONS

We have introduced a new nonrigid tracking and unsupervised modelling approach with both global and local properties of the image domain. This nonparametric learning makes no assumption about the global structure of the hand gestures thus, no background modelling nor training sets are required. Due to the number of features A-GNG uses, the topological relations are preserved and nodes correspondences are retained between tracked configurations. The proposed approach is robust to object transformations, and can prevent fold-overs of the network. The model is learned automatically by tracking the nodes and evaluating their position over a sequence of k frames. The algorithm is com-

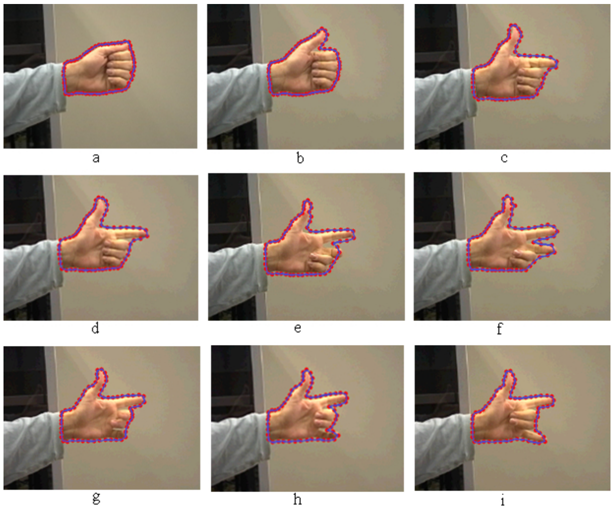

Figure 11: Tracking a gesture. The images correspond from left to right and from top to bottom to every 10th frame of a 90 frame sequence. In each image the red points indicate the nodes and their adaptation after 4 iterations.

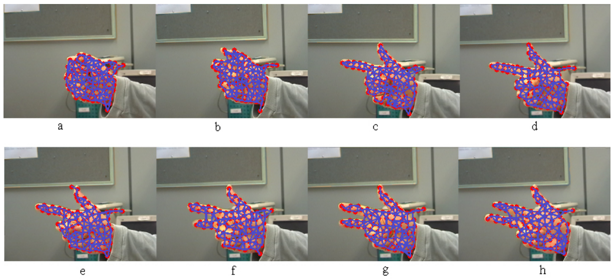

Figure 12: Tracking a gesture in a cluttered background. The images correspond from left to right and from top to bottom to every 10th frame of a 90 frame sequence.

putationally inexpensive, it preserves topological relations based on global and local transformations, thus network is partially supervised, and it can track in an unsupervised manner $2D$ hand gestures in cluttered backgrounds using skin colour information.

6. REFERENCES

[1] A. Angelopoulou, A. Psarrou, G. Gupta, and J. García. Nonparametric Modelling and Tracking with Active-GNG. *IEEE Workshop on Human Computer Interaction, HCI 2007, in conjuction with ICCV 2007, LNCS 4796*, pages 98–107, 2007.

[2] H. Bauer and K. Pawelzik. Quantifying the neighbourhood preservation of self-organizing feature maps. *IEEE Transactions on Neural Networks*, 3(4):570–579, 1992.

[3] A. Baumberg and D. Hogg. Learning Flexible Models from Image Sequences. *In Proc. of the* 3^{rd} *European Conference on Computer Vision*, 1:299–308, 1994.

[4] C. M. Bishop. Neural networks for pattern recognition. *Oxford Uni. Press*, 1995.

[5] T. F. Cootes, C. J. Taylor, D. H. Cooper, and G. J. Active Shape Models - Their Training and Application. *Comp. Vision Image Underst.*, 61(1):38–59, 1995.

[6] D. Cremers, T. Kohlberger, and C. Schnorr. Shape statistics in kernel space for variational image segmentation. *Pattern Recognition*, 36(9):1929–1943, 2003.

[7] H. R. Davies, J. C. Twining, F. T. Cootes, C. J. Waterton, and J. C. Taylor. A Minimum Description Length Approach to Statistical Shape Modeling. *IEEE Transaction on Medical Imaging*, 21(5):525–537, 2002.

[8] N. Duta, A. K. Jain, and M.-P. Dubuisson-Jolly. Automatic construction of 2D shape models. *IEEE Trans. PAMI*, 23(5):433–446, 2001.

[9] F. Florez-Revuelta, J. M. Garcia-Chamizo, J. Garcia-Rodriguez, and A. Hernandez-Saez. Geodesic topographic product: An improvement to measure topology preservation of self-organizing neural networks. *Advances in Artificial Intelligence-IBERAMIA 2004*, 3315:841–850, 2004.

[10] B. Fritzke. A growing Neural Gas Network Learns Topologies. *In Advances in Neural Information Processing Systems 7 (NIPS'94)*, pages 625–632, 1995.

[11] S. Furao and O. Hasegawa. An incremental network for on-line unsupervised classification and topology learning. *The Journal of Neural Networks 19*, pages 90–106, 2005.

[12] A. Hill, C. Taylor, and A. Brett. A Framework for Automatic Landmark Identification Using a New Method of Nonrigid Correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3):241–251, 2000.

[13] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active Contour Models. *In Proc. of the* 1^{st} *Internationl Conference on Computer Vision, IEEE Computer Society Press*, pages 259–268, 1987.

[14] S. Lankton, D. Nain, A. Yezzi, and A. Tannenbaum. Hybrid geodesic region-based curve evolutions for image segmentation. *In Proc. SPIE: Med. Img.*, 6510:65104U, 2007.

[15] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface quality hand tracking. *In Proc. of the* 6^{th} *European Conference on Computer Vision, ECCV 2000*, 2:3–19, 2000.

[16] C. Nastar and N. Ayache. Fast segmentation, tracking and analysis of deformable objects. *In Proc. of the* 4^{th} *International Conference on Computer Vision, ICCV'93*, pages 275–279, 1993.

[17] L. H. Staib and J. S. Duncan. Parametrically deformable contour models. *In IEEE conference on Computer Vision and Pattern Recognition*, pages 427–430, 1989.

[18] E. B. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky. Distributed Occlusion Reasoning for Tracking with Nonparametric Belief Propagation. *In Advances in Neural Information Processing Systems*, 17:1369–1376, 2005.

[19] A. L. Yuille, D. S. Cohen, and P. Hallinan. Feature extraction from faces using deformable templates. *Int. J. Computer Vision*, (8):99–112, 1992.