Adaptive Learning in Motion Analysis with Self-Organising Maps

Anastassia Angelopoulou, José García-Rodríguez, Alexandra Psarrou, Gaurav Gupta, and Markos Mentzelopoulos

Abstract—Growing models have been widely used for clustering or topology learning. Traditionally these models work on stationary environments, grow incrementally and adapt their nodes to a given distribution based on global parameters. In this paper, we present an enhanced unsupervised self-organising network for the modelling of visual objects. We first develop a framework for building non-rigid shapes using the growth mechanism of the self-organising maps, and then we define an optimal number of nodes without overfitting or underfitting the network based on the knowledge obtained from information-theoretic considerations. This model is used to the representation of motion in image sequences by initialising a suitable segmentation. We present experimental results for hands and we quantitatively evaluate the matching capabilities of the proposed method with the topographic product.

I. INTRODUCTION

In recent years, self-organised models have gained popularity in applications related to computer vision, man-machine interaction, and biometric systems like: segmentation and representation of objects [10], [18], [17], [31], [34], tracking objects [6], [11], recognition of gestures [4], [19], biomedicine [2], [8], and 3D reconstruction [1], [7], [21], [23], [29].

Many researchers have modified growing models in applications related to computer vision, man-machine interaction, and biometric systems. Fritzke [15] presented variations of the original GNG [14] algorithm to deal with non-stationary distributions, which he called the Growing Neural Gas with Utility (GNG-U), and a semi-supervised variation SNG [12] combined with Radial basis function (RBF) networks. Marsland et al. [24] present a variation, named Grow When Required (GWR) algorithm able to add nodes whenever the network does not sufficiently match the input. Furao and Hasegawa [16] introduced an incremental learning GNG model to handle online non-stationary problems. Prudent and Ennaji [27] also proposed an incremental learning model based on the Adaptive Resonance Theory (ART) mechanism, called the Incremental GNG (IGNG). A very promising method but for only supervised or semi-supervised learning and with applications in low-dimensional data visualisation input space. Qin and Suganthan [28] proposed the Robust GNG (RGNG) algorithm for unsupervised clustering. The algorithm added several techniques to the original GNG algorithm to reduce the sensitivity of the algorithm to prototype initialization, input sequence, and outliers. Angelopoulou et al. [3] also used the GNG to automatically obtain interest points in medical shapes and built statistical shape models. Frezza-Buet [11] has slightly modified the original GNG algorithm, called GNG-T, by continuously performing vector quantisation over a distribution that changes over time. Wu et al. [34] and Stergiopoulou et al. [32] suggest the use of self-organising networks for human-machine interaction. Xiang Cao et al. [6] and Vasquez et al. [33] propose amendments to self-organising models for the characterisation of the movement.

A disadvantage of these approaches is that there is no consideration of time constraints or exploitation of knowledge gained from previous frames. Having a sequence of images this can be avoided by using an incremental model. One of the most important characteristics of the GNG is that it does not require the restarting of the initialisation of the network for every image in a sequence of k frames.

In this paper, we are interested in the minimisation of the user intervention in the learning process of the network by utilising an automatic criterion for maximum node growth in the GNG network based on different parameters. We achieve that by taking into consideration that human skin has a relatively unique colour and the complexity or simplicity of the proposed model is decided by information-theoretic measures. The remainder of the paper is organised as follows. Section 2 introduces the framework for object modelling using topological relations. Section 3 proposes an approach to minimise the user intervention in the termination of the network using knowledge obtained from information-theoretic considerations. In Section 4 our method is applied to real and artificial shapes before conclusions are drawn in Section 5.

II. CHARACTERISING 2D OBJECTS USING GNG

GNG [14] is an unsupervised incremental self-organising network independent of the topology of the input distribution or space. It uses a growth mechanism inherited from the
Growth Cell Structure [13] together with the Competitive Hebbian Learning (CHL) rule [25] to construct a network of the input data set. In GNG, the growing process starts with two nodes, and new nodes are incrementally inserted until a predefined condition is satisfied, such as the maximum number of nodes or available time. 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 and new connections between the winner node and its topological neighbours are created.

Identifying the points of the image that belong to objects allows the network to obtain an induced Delaunay triangulation of the objects. Let an object $O = [O_G, O_A]$ be defined by its geometry and its appearance. The geometry provides a mathematical description of the object’s shape, size, and parameters such as translation, rotation, and scale. The appearance defines a set of object’s characteristics such as colour, texture, and other attributes.

Given a domain $S \subseteq \mathbb{R}^2$, an image intensity function $I(x, y) \in \mathbb{R}$ such that $I : S \rightarrow [0, I_{\max}]$, and an object $O$, its standard potential field $\Psi_T(x, y) = f_T(I(x, y))$ is the transformation $\Psi_T : S \rightarrow [0, 1]$ which associates to each point $(x, y) \in S$ the degree of compliance with the visual property $T$ of the object $O$ by its associated intensity $I(x, y)$.

The likelihood of the required number of nodes to describe the topology of an image $y$ is:

$$ p(y|x) = \prod_{u \in \Omega(x)} p_{\text{skin}}(u) \prod_{v \in \Gamma \setminus \Omega(x)} p_{\text{bkgd}}(v) $$

and $e_T = \prod_{u \in \Omega(x)} p_{\text{skin}}(u) + \prod_{v \in \Gamma \setminus \Omega(x)} p_{\text{bkgd}}(v)$. $e_T$ is a similarity threshold and defines the accuracy of the map. If $e_T$ is low the topology preservation is lost and more nodes need to be added. On the contrary, if $e_T$ is too big then nodes have to be removed so that Voronoi cells become wider. For example, let us consider an extreme case where the total size of the image is $I = 100$ pixels and only one pixel represents the object of interest. Let us suppose that we use $e_T = 100$ then the object can be represented by one node. In the case where $e_T \geq I$ then overfitting occurs since twice as many nodes are provided. In our experiments the numerical value of $e_T$ ranges from $100 \leq e_T \leq 900$ and the accuracy depends on the size of the objects’ distribution. The difference between choosing manually the maximum number of nodes and selecting $e_T$ as the similarity threshold, is the preservation of the object independently of scaling operations.

The following are the steps of the automatic criterion:

1. **The input distribution as the set of points in the image:**
   $$ A = S $$
   (1)

2. **The probability density function according to the standard potential field obtained for each point of the image:**
   $$ p(x, y) = \Psi_T(x, y) $$
   (2)

3. **Learning takes place with our modified GNG algorithm [22]. So, during this process, the neural network is obtained which preserves the topology of the object $O$ from a certain feature $T$. Therefore, from the visual appearance $O_A$ of the object is obtained an approximation to its geometric appearance $O_G$.**

Fig. 1. First shape of each of the first 10 objects in coil-100, showing the original image, the thresholded region, and the modified GNG contour representation.
added to the modified GNG algorithm to minimise user intervention in the learning process.

**Algorithm Stopping Criterion for Maximum Node Growth**

**Input:** Segmented pixels \( \Omega(x) \) from an unknown image \( I \)

**Output:** likelihood \( p(y|x) \), TPG

1) Initialise \( x_e \) the network parameter to: 
2) Obtain \( e_T \) value \( \leq \prod_{u \in \Omega(x)} p_{\text{skin}}(u) + \prod_{v \in \mathcal{X} \setminus \Omega(x)} p_{\text{bkgd}}(v) \)
3) for every pixel \( x \) do 
   if \( 100 \leq e_T < 500 \) or \( e_T == 900 \) then
   Use Equation 4 to find \( x_e \).
   if Number of prototypes \( x_e \leq 50 \) then
   Increment \( e_T \) until \( x_e \geq 120 \).
   end if
   if Number of prototypes \( x_e \geq 200 \) then
   Decrement \( e_T \) until \( 120 \leq x_e < 200 \).
   end if
   else
   \( x_e \) = maximum prototypes, update list, update connections.
   end if
4) end for

We can describe the optimum number of similarity thresholds, required for the accuracy of the map for different objects, as the unknown clusters \( K \), and the network parameters as the mixture coefficients \( W_K \), with \( d \)-dimensional means and covariances \( \Theta_K \). To do that, we use a heuristic criterion from statistics known as the Minimum Description Length (MDL) [30], which does not require an estimation of the complexity of \( p(Y) \) as it is in the case of the conditional entropy heuristic criterion [5]. The MDL criterion takes the general form of a prediction error, which consists of the difference of two terms:

\[
E = \text{model likelihood} - \text{complexity term} \tag{5}
\]

a likelihood term that measures the model fit and increases with the number of clusters, and a complexity term, used as penalty, that grows with the number of free parameters in the model. Thus, if the number of clusters is small we get a low value for the criterion because the model fit is low, while if the number of clusters is large we get a low value because the complexity term is large.

The information-criterion MDL of Rissanen [30], is defined as:

\[
\text{MDL}(K) = -\ln[L(X|W_K, \Theta_K)] + \frac{1}{2} M \ln(N) \tag{6}
\]

where

\[
L(X|W_K, \Theta_K) = \max_{i=1}^{N} p(x_i|W_K, \Theta_K) \tag{7}
\]

The first term \( -\ln[L(X|W_K, \Theta_K)] \) measures the model probability with respect to the model parameter \( W_K, \Theta_K \) defined for a Gaussian mixture by the mixture coefficients \( W_K \) and \( d \)-dimensional means and covariances \( \Theta_K \). The second term \( \frac{1}{2} M \ln(N) \) measures the number of free parameters needed to encode the model and serves as a penalty for models that are too complex. \( M \) describes the number of free parameters and is given for a Gaussian mixture by \( M = 2dK + (K - 1) \) for \( (K - 1) \) adjustable mixture weights and \( 2d \) parameters for \( d \)-dimensional means and diagonal covariance matrices.

The optimal number of similarity thresholds can be determined by applying the following iterative procedure:

- For all \( K, (K_{\min} < K < K_{\max}) \)
  - (a) Maximize the likelihood \( L(X|W_K, \Theta_K) \) using the EM algorithm to cluster the nodes based on the similarity thresholds applied to the data set.
  - (b) Calculate the value of MDL(K) according to Equations 6 and 7.
- Select the model parameters \( (W_K, \Theta_K) \) that corresponds to minimisation of the MDL(K) value.

Figure 2 shows the value of MDL(K) for clusters within the range of \((1 < K < 18)\). We have doubled the range in the MDL(K) minimum and maximum values so we can represent the extreme cases of 1 cluster which represents the whole data set, and 18 clusters which over classify the distribution and corresponds to the overfitting of the network with similarity threshold \( e_T = 900 \). A global minimum and therefore optimal number of clusters can be determined for \( K = 9 \) which indicates that the best similarity threshold that defines the accuracy of the map without overfitting or underfitting the data set is \( e_T = 500 \). To account for susceptibility for the EM cluster centres as part of the MDL(K) initialisation of the mixture coefficients the measure is averaged over 10 runs and the minimal value for each configuration is selected.

The procedure is summarised in the following steps.

**Algorithm MDL(K) Value**

**Input:** TPG

**Output:** MDL(K)

1) Initialise TPG (Previous Algorithm)

2) while current number of prototypes = \( x_e \) do
   Calculate MDL(K) according to Equations 6 and 7.
   Save position of all prototypes and average MDL(K) over 10 runs.

3) end while

We can now use this optimal network to track objects locally wherever common regions are found. To do that shape information and colour information from the 1st and any subsequent frames are added to the TPG map and can be used for the learning in a sequence of \( k \) frames. The segmented frame and the stored shape and colour information
in each node is given by:

\[ S(x; P(g(x,y); t) = p(k|x) \propto P(g(x,y), t - 1), TPG_t-1) \tag{8} \]

Figure 3 shows the convergence of the network with shape and posterior probability per node.

IV. EXPERIMENTS

We tested our modified GNG network on a data set of hand images recorded from 5 participants each performing eight gestures (Figure 4) that frequently appear in sign language. To create this data set we have recorded images over several days and a simple webcam was used with image resolution 800 × 600. In total, we have recorded over 7500 frames, and for computational efficiency, we have resized the images from each set to 300 × 225, 200 × 160, 198 × 234, and 124 × 123 pixels. We obtained the data set from the University of Alicante, Spain and the University of Westminster, UK. Also, we tested our method with 49 images from Mikkel B. Stegmann\(^1\) online data set. In total we have run the experiments on a data set of 174 images. Since the background is unambiguous the network adapts without occlusion reasoning. For our experiments only complete gesture sequences are included. There are no gestures with partial or complete occluded regions, which means that we do not model multiple objects that interact with the background. Furthermore, we have performed the experiments having in mind specific multiple objects that interact with the background. The quality and stability of the results at close range makes it worthwhile for webcam or green screen sign language applications which share a close range viewing distance and a relatively uncluttered background.

To classify a region as a hand or face we take into account domain knowledge information that respects always some proportions found in hands and human faces [20]. To do that we find the centroid, height and width of the connected nodes in the networks as well as the percentage of skin in the rectangular area (Figure 5). Since the height to width ratio for hands and human faces fall into a small range, we are able to reject or accept if the topology of a network represents or not a hand. Studies [9], [20], have shown that the height to width ratio of human face and hands fall within a range defined based on the well known Golden Ratio (Equation 5.6). Thus, we consider a network as a hand or not if the height to width ratio of the region falls within a range of the Golden Ratio ± Tolerance. In the case where the hand is in a folded posture the rule still applies but with different percentage for the Tolerance. The values for the Tolerance were found by experimentation, and range from 0.5 to 0.7 based on the hand posture.

\[ \phi \equiv \frac{\text{height}}{\text{width}} \equiv \frac{(1 + \sqrt{5})}{2} \tag{9} \]

Table I shows topology preservation, execution time, and number of nodes when different variants in the \(\lambda\) and the \(K\) are applied in a hand as the input space. Faster variants get worse topology preservation but the network converges quickly. However, the representation is sufficient and can be used in situations where minimum time is required like

\(^{1}\text{http://www2.imm.dtu.dk/~aam/}\)
A step is performed globally even if $a_{\text{active}} > 900$. $e_4$ is set to reflect $\lambda = 500$ nodes. $T = 900 = 100 = 100 V = 500$ and $\times T$ (left image) and the times less compared to the snake. $(\text{right image})$ pixels with a network map of nodes to the same object (class probability of pixels $- e$, knowledge of the domain, which means it can deform Fig. 5. Example of correctly detected hands and face based on the golden ratio regardless of the scale and the position of the hands and the face. (a) original image, (b) after applying EM to segment skin region, and (c) hand and face detector taking into account the connected nodes in the networks as well as the percentage of skin in the rectangular area.

online learning for detecting obstacles in robotics where you can obtain a rough representation of the object of interest in a given time and with minimum quality.

Figure 6 shows the plots of the MDL($K$) values versus the number of clusters for minimum, maximum and approximate match similarity threshold ($e_T = 100$, $e_T = 900$, and $e_T = 500$). As the similarity threshold increases the optimum number for the MDL($K$) values increases as well with an optimum growth at $K = 9$.

![Image](image_url)

Fig. 6. (a), (b), and (c) Plot of data set for similarity thresholds $e_T = 100$, $e_T = 500$, and $e_T = 900$. (d), (e), and (f) Plot of the MDL($K$) values versus the number of clusters centres generated by the similarity thresholds during the growth process of the GNG.

Table II shows the topographic product for a number of nodes. We can see that the insertion of more nodes makes no difference to the object’s topology. Based on the maximum size of the network an optimum result is achieved when at least half of the network is developed. Table II shows that for the different type of gestures this optimum number is in the range $> 90$ and $< 130$.

Figure 7 presents the adaptation with a network map of 72 nodes to the same object (class probability of pixels belonging to the objects of interest $P(O)$ is set to reflect the size of objects in an image and be $1 - P(B)$ (the background probability) but with different image size. The image has been re-scaled by half the size of the original image resolution. In both cases the adaptation is correct and the topology is preserved independently of the scaling of the image.

Our modified GNG network has also been compared to the methodology of the active snake model. The snake converges when all the forces achieve an equilibrium state. The drawbacks with this method are that the snake has no $a$ priori knowledge of the domain, which means it can deform to match any contour; this attribute is not desirable if we want to keep the specificity of the model or preserve the physical attributes such as geometry, topological relations, etc., and that the active step is performed globally even if parts of the snake have already converged. Figure 8 shows the tracking of a hand gesture using the modified GNG tracker. Figure 9 shows the fitting results of a snake applied to the same gesture. Figure 9 (a) is the original state of the snake after manually locating an area around the hand. The closer we allocate landmark points around the hand the faster the convergence of the snake. The snake after a number of iterations converges to the palm of the hand but fails to convergence around the thumb. The parameters for the snake are summarised in Table III. The execution time for modified GNG is approximately 4 times less compared to the snake.
Fig. 8. Tracking a gesture. The images correspond from left to right and from top to bottom to every 10th frame of a 190 frame sequence. In each image the red points indicate the nodes and their adaptation after 4 iterations.

The computational and convergence results are summarised in Table IV.

<table>
<thead>
<tr>
<th>Hand Constants</th>
<th>Iterations</th>
<th>Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence (a)</td>
<td>α = 0.110</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>β = 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>γ = 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>κ = 0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D_{min} = 0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D_{max} = 2</td>
<td></td>
</tr>
<tr>
<td>Sequence (b)</td>
<td>α = 4</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>β = 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>γ = 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>κ = 0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D_{min} = 0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D_{max} = 2</td>
<td></td>
</tr>
<tr>
<td>Sequence (c)</td>
<td>α = 4</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>β = 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>γ = 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>κ = 0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D_{min} = 0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D_{max} = 2</td>
<td></td>
</tr>
<tr>
<td>Sequence (d)</td>
<td>α = 4</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>β = 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>γ = 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>κ = 0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D_{min} = 0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>D_{max} = 2</td>
<td></td>
</tr>
</tbody>
</table>

Table IV

Convergence and Execution Time Results of Modified GNG and Snake

<table>
<thead>
<tr>
<th>Method</th>
<th>Convergence (Iteration times)</th>
<th>Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>snake</td>
<td>20</td>
<td>5.60</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>12.01</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>15.20</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>15.29</td>
</tr>
<tr>
<td>modified GNG</td>
<td>2</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2.17</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4.88</td>
</tr>
</tbody>
</table>

In all our experiments, the parameters of the network are as follows: \( \lambda = 100 \) to 1000, \( \epsilon_x = 0.1 \), \( \epsilon_n = 0.0005 \), \( \Delta x_1 = 0.5 \), \( \Delta x_i = 0.0005 \), \( \alpha_{max} = 125 \). For the MDL(K) value we have experimented with cluster centres within the range of \( 1 < K < 18 \).

V. Conclusions and Future Work

Based on the capabilities of GNG to readjust to new input patterns without restarting the learning process, we developed an approach to minimise the user intervention by utilising an automatic criterion for maximum node growth. This automatic criterion for GNG is based on the object’s distribution and the similarity threshold (\( e_T \)) which determines the preservation of the topology. The model is then used to the representation of motion in image sequences by initialising a suitable segmentation. During testing we found that for different shapes there exists an optimum number that maximises topology learning versus adaptation time and MSE. This optimal number uses knowledge obtained from information-theoretic considerations.

REFERENCES


