MATCHING FLAT OBJECTS IN 3D SPACE USING DYNAMIC LINK ARCHITECTURE

H.C. Sim and R.I. Damper

Image, Speech and Intelligent Systems (ISIS) Research Group,
Department of Electronics and Computer Science,
University of Southampton,
Southampton SO17 1BJ, UK.
(email: {hcs95r|rid}@ecs.soton.ac.uk)

ABSTRACT

This paper describes a novel approach for planar object matching which is invariant to 3D perspective transformation and partial occlusion. The object's domain is not restricted to purely 2D items; it includes fairly flat real objects such as a pair of scissors. The proposed system uses multi-view model representations and objects are recognised by self-organised dynamic link matching. The merit of this approach is that it offers a compact framework for concurrent assessments of multiple match hypotheses by promoting competitions and/or co-operations among several local mappings of model and test image feature correspondences. Our experiments show that the system is very successful in recognising objects deformed because of perspective distortion, even in rather cluttered scenes.

1. INTRODUCTION

To date, most existing object matching systems are only useful in recognizing 2D objects which are subjected similarity transformations. By contrast, this paper takes a more generalized approach for planar object matching which is invariant to 3D perspective transformation and partial occlusion.

Our approach is inspired by dynamic link architecture (DLA) [1], which appears to be one of the most versatile and realistic neural-technique for arbitrary object matching which was proposed in [1, 2, 3]. Dynamic link matching involves 2 layers of neuronal maps labeled $\mathcal{I}$ (representing image graph) and $\mathcal{M}$ (representing model graph). Each neuron represents a local feature detector. There are intra-layer connections which encode geometric relations among neighboring neurons within a layer, and inter-layer dynamic links connecting neurons on different layers. On the basis of local feature similarity, the system rapidly modifies link weights in a Hebbian fashion to establish neighborhood-preserving mappings which connect pairs of points with similar local features. In brief, the system undergoes unsupervised self-organization of dynamic links with the aid of appropriate intra- and inter-layer dynamics.

It is difficult to use the original DLA approach to solve our problems effectively without special modifications. There are two fundamental problems here: 1) first, it is difficult to determine local feature similarity given the huge amount perspective distortions we try to accommodate (compared with [2] and [1]) resulting in a huge variations of local features; 2) placement of image neuronal map $\mathcal{I}$ is far more complicated and demanding in our problems because of cluttered test image, partial occlusion and severe perspective distortions. To overcome the first problem, a self-organized hierarchical indexing system is used to cluster highly correlated model features so that an unknown test image feature only needs to be compared with

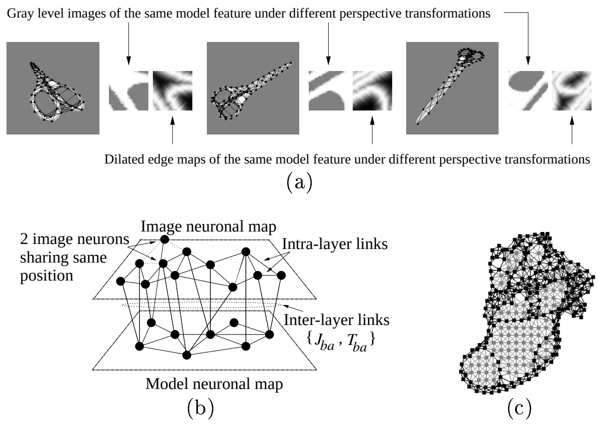


Figure 1: (a) Varying views of model features. (b) Modified dynamic link Architecture. (c) Model neuronal map with primary features (black dots) and secondary features (gray dots).

a small number of most likely matching model features. As for the second problem, a novel image neuronal map construction process is devised to generate better aligned feature correspondences with less redundancies among interlayer attributes. As a result of these new features, special adaptations to the original DLA are needed in order to fully exploit the potential of the proposed system.

Following these introductory paragraphs, section 2 describes the model representation and feature extraction modules. Section 3 details the hierarchical indexing system. Section 4 describes the minimization process used to align matching image and model feature pairs. The modified dynamic link matching process is highlighted in section 5. Some experiments are reported in section 6. Finally, concluding remarks and future work are presented in section 7.

2. MODEL REPRESENTATION

Unlike conventional DLA where image neurons have interlayer links to all model neurons, an image neuron $a \in \mathcal{I}$ in the modified DLA has only one inter-layer connection, but many inter-layer connections may converge at the same model neuron (cf. Figure 1(b)). Several image neurons with dissimilar matching counterparts in the model neuronal map may occupy the same position in the test image. Such unusual connectivity between neurons is chosen because closely connected model neurons may be transformed into the same co-ordinates under perspective transformations.

The features are extracted by convolving the 2D image

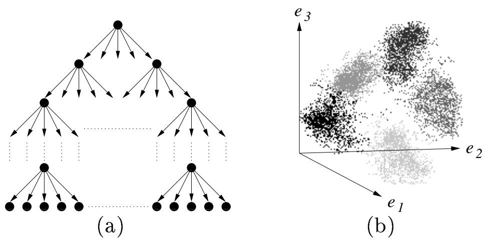


Figure 2: (a) Tree-like self-organized hierarchical indexing system. (b) 5 Clusters formed at a root node.

$I(\vec{x})$ or with a set of 2D Gabor functions [1, 3] given by:

$$\psi_{\vec{k}}(\vec{x}) = \frac{\vec{k}^2}{\sigma^2} \exp\left(-\frac{\vec{k}^2\vec{x}^2}{2\sigma^2}\right) \left[\exp\left(-i\vec{k}\vec{x}\right) - \exp\left(\frac{\sigma^{-2}}{2}\right)\right], \tag{1}$$

where $\vec{k} = \{w_\nu \cos(\theta_\mu), w_\nu \sin(\theta_\mu)\}$ determines the center frequencies w_ν and orientations θ_μ of the 2D Gabor filters. The system utilizes 2 different frequency levels i.e. $w_\nu = \{0.27, 0.58\}$ and 8 different orientations i.e. $\theta_\mu = \frac{\pi\mu}{8}$, $\mu \in \{0, \ldots, 7\}$. The relative bandwidth σ is set at 2.0. Gabor wavelet representations may not be useful for many artificial objects with homogeneous surfaces. To overcome this, Gabor filters are applied to both the gray-level intensity image and the edge map as shown in Figure 1(a). The former is still useful for highly structured surface markings. The edge map is dilated by passing a Gaussian filter along the edge pixels which helps to smooth the contour and join up broken ones. A multi-scale strategy is used to provide scale invariance between scaling factors of 0.5 and 1.5.

A model neuronal map $\mathcal{M}$ is essentially a connected grid of local features or neurons as illustrated in Figure 1(c). Each dot represents a local feature and arcs represent intra-layer connections. A multi-view approach is used to capture Gabor filter responses of each model neuron when the object is seen from 140 evenly spaced viewpoints.

3. HIERARCHICAL INDEXING SYSTEM

A tree-like self-organized hierarchical indexing system as shown in Figure 2 is created to cluster model features with almost identical Gabor responses. In this way, a test image feature only needs to be validates against a few clusters of most likely matching model features. There are 2 types of model neurons – *primary* or *secondary* features (cf. Figure 2(a)). Primary features (black dots in Figure 1(c)) are sample points along object boundaries and user-selected surface contours. Secondary ones (gray dots) are automatically spaced foreground sample points. To reduce storage space requirements, only low frequency filter responses of primary features are stored. Secondary features play the auxiliary role of supporting local matching of primary features. At each node of the indexing tree, sampled Gabor filter responses are grouped as follows: 1) discrete Karhunen-Loéve expansion [4] is applied to extract effective features among Gabor responses; 2) unsupervised clustering is used to form highly correlated groups of principal features produced in step 1; 3) finally, Fisher's transform [4] is applied to extract the optimal set of features which is most effective for preserving class separability. The aforesaid unsupervised clustering is an iterative k-means approach where model feature is attracted to the nearest cluster according to its normalized Mahalanobis distance to each cluster. The initial k centers needed by k-means algorithm are determined using vector quantization based on "neural-gas" networks [5]. 5 clusters or less are formed at each node in the tree, until groups of approximately 200 Gabor filter responses are formed at the leaves. The deepest tree encountered in our experiments has only 5 levels. Figure 2(c) shows 5 distinct clusters created at the root of a tree used in Experiment 3. When an unknown image Gabor response is presented, the test feature undergoes appropriate canonical space projections at each node as it traverses down the tree following paths within 1.3 times the shortest Mahalanobis distance determined at each encountered tree node. Two indexing trees are created, one for Gabor responses of pixel intensity and the other for dilated edge map (cf. section 2).

4. FEATURE MATCHING AND ALIGNMENT

Upon reaching the leaves, edge gradient and principal orientation of gray level distributions are used to prune away more unlikely matches, both attributes must be $\leq 20°$. On average, a given image feature only needs to match against ≤ 80 model features at the leaves. Similarity between model feature $b \in \mathcal{M}$ and test image feature $a \in \mathcal{I}$ is computed as follows:

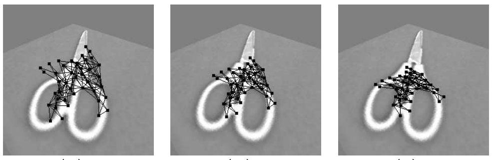
$$\mathcal{S}(P_{ba}, \vec{b}, \vec{a}) = \frac{\vec{b} \cdot \vec{a}}{\|\vec{b}\|\|\vec{a}\|} \min\left(\frac{\|\vec{b}\|}{\|\vec{a}\|}, \frac{\|\vec{a}\|}{\|\vec{b}\|}\right), \tag{2}$$

where $\mathcal{S}(\cdot)$ is the similarity function. $\vec{b}$ and $\vec{a}$ are Gabor filter responses of model and test image features respectively. $\|\cdot\|$ denotes the magnitude of a vector. P_{ba} denotes the perspective transformation, i.e. $\vec{b}$ is the Gabor filter responses of model sample point b when the object undergoes perspective transformation given by P_{ba}. All model and image feature pairs which yield a value of ≥ 0.75 in equation (2) for Gabor filter responses of pixel intensity or dilated edge map are deemed possible matches. This group of highly probable feature correspondences is further processed using high frequency Gabor response with equation (2). In brief, high frequency Gabor filters are used to measure the local gray-level similarity within narrow bands of foreground surrounding the dilated model and test image edge contours.

Fine tuning of feature alignment is needed to provide acceptable pose estimation. There are 7 perspective transformation parameters that must be optimized, i.e. P_{ba} = elevation angle (ϕ_{ev}), azimuth angle (ϕ_{az}), translation (d_x, d_y), zoom factor s, and view reference point co-ordinates (v_x, v_y). The cost function is formulated as:

$$\mathcal{C}(P_{ba}) = \frac{1}{N} \sum \left(1 - \exp\left(-\alpha D(\vec{x})\right) |\cos(\triangle(\vec{x}))|\right), \tag{3}$$

where $D(\vec{x})$ is the distance between a superimposed model edge point to the nearest image edge point, $\triangle(\vec{x})$ is the gradient difference between model and image edge points, α is the smoothing factor and N is the number of model edge points. Gabor wavelet transforms is not used here because they are insensitive to small changes in transformation parameters and are expensive to calculate. Simulated annealing with downhill simplex minimization is used to align feature correspondences quickly. There are 8 vertices on the simplex given 7 input parameters. These vertices are initialized to the implied perspective transformations of the 8 most likely feature correspondences according to equation (2). Hence, competition among hypothesized matches can still occur at this level. Following this rough optimization, conjugate gradient minimization is used to refine the resultant transformation parameters. Let the ordered pair $\{b \in \mathcal{M}, a \in \mathcal{I}\}$ denotes a feature correspondence considered by the alignment process, and L denotes the levels of model neurons radiating from b, where $L = 0$ refers to b itself, $L = 1$ refers to model neurons with direct intra-layer links to b and so on. Only edges encompassed within area surrounding L levels of projected model neurons in the test image are used for feature alignment, e.g. Figure 3 shows



(a) (b) (c)

Figure 3: Progressive feature alignment from initial placement (a) to final alignment (c).

the progressive alignment of feature correspondences for $L = 3$ levels of model neurons. Note that both primary and secondary features are used here.

5. DYNAMIC LINK MATCHING

At the heart of the proposed system, a central matching unit based on the modified DLA draws evidences of local model and test feature correspondences from supporting modules described in the preceding sections to formulate a suitable match. The matching phase is divided into 2 stages: in the first stage, neuronal map $\mathcal{I}$ is created; in the second stage, our modified dynamic link matching process is applied to discern the desired match.

5.1. Stage One

Edge pixels are set activated within square window denoted by $\mathcal{W}$ which is randomly positioned in test image. Matching counterparts are found for activated edge pixels. Only edge pixels are activated because primary features stored in the indexing system (cf. sections 3) are Gabor filter responses of areas surrounding model edge pixels. Simulated annealing and conjugate gradient minimization (cf. section 4) are used to resolve the competition between conflicting hypotheses and to align local features. New image neurons $a \in \mathcal{I}$ are created for L levels of model neurons used in the alignment process. If $g(a^*) \in \mathcal{M}$ denotes the model counterpart of image neuron $a^* \in \mathcal{I}$, then similarity measures $T_{ba} = \{T'_{ba}, T''_{ba}\}$ are assigned to the newly created feature correspondence $\{b \in \mathcal{M}, a \in \mathcal{I}\}$ as follows:

$$\begin{aligned} T'_{ba} &= \mathcal{S}(P_{ba}, \vec{b}, \vec{a}) \\ T''_{ba} &= \lambda T'_{ba} + (1-\lambda)\mathcal{H}(b, a), \end{aligned} \tag{4}$$

$$\mathcal{H}(b,a) = \frac{1}{2|\Lambda_b|} \sum_{b^* \in \Lambda_b} \max_{\substack{a^* \in \Lambda_a \\ g(a^*) = b^*}} \left[\mathcal{S}(P_{ba}, \vec{b^*}, \vec{a}^*) + \vec{\mathcal{X}}_{aa^*} \right] \tag{5}$$

where $\mathcal{S}(\cdot)$ is given in (2). $\mathcal{H}_(b,a)$ in the second part of T''_{ba} measures features conformity within the neighborhood of b and a, where where Λ_k is the set of neurons directly connected to neuron k via intra-layer links and $|\Lambda_k|$ denotes the number of neurons in Λ_k. $\vec{\mathcal{X}}_{aa^*} \in [0,1]$ denotes the difference between ideal and actual displacements between image neurons a and a^*. $\lambda = 0.7$ dictates the relative significance of the two terms. Obviously, T'_{ba} applies to the outermost level L of model neurons used in the alignment process, whereas T''_{ba} applies to model neurons from levels 0 to $L-1$.

A control mechanism based on competition among conflicting hypotheses is used to prevent proliferation of spurious feature correspondence. Forbidden zones are introduced whereby no two image neurons $a \in \mathcal{I}$ are allowed to converge at the same model neuron $b \in \mathcal{M}$. There are 2 types of forbidden zones namely, fixed and floating. Each 256×256 test image is divided into 64 non-overlapping 32×32 fixed forbidden zones $\mathcal{R}_i, i = 1 \ldots 64$ as shown in Figure 4. Floating forbidden zones are 16×16 windows defined around new added image neurons $a \in \mathcal{I}$ where the uniqueness constraint must also apply (only one is shown in Figure 4). If conflicts occur where ≥ 2 image neurons converge at the same model neuron in these forbidden zones, then the one with highest Gabor response similarity given by (2) is retained, while the rest are destroyed. Intra-layer links connected to the expunged image neurons are reconnected to the selected image neurons. An example of this process is illustrated in Figure 4. T_{ba} of affected feature correspondences which have modified intra-layer connections are recalculated using (4). Under this scheme, the number of image neurons will not exceed $64 \times |\mathcal{M}|$, where $|\mathcal{M}|$ is the number of model neurons.

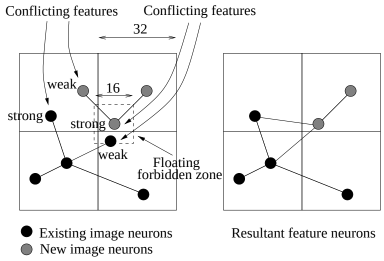


Figure 4: Competition among conflicting feature correspondences.

The size of activation windows $\mathcal{W}$ are set at 32×32. $L = 3$ levels of model neurons are used in the local alignment process.

5.2. Stage Two

Inter-layer dynamic links $J_{ba} \in [0,1]$ are established to connect pairs of potentially matching model neuron b and image neuron a. J_{ba} is only created for feature correspondences with T''_{ba}, i.e.

$$J_{ba} = \frac{T''_{ba}}{\sum_{a'} T''_{ba'}}, \tag{6}$$

where a' are image neurons connected to b. Image blob denoted by $\mathcal{B}(\vec{x}_c)$ is randomly activated as part of system dynamics. Unlike classical DLA, image blobs used here are simply square windows centered at randomly selected test image point $\vec{x}_c$. Effective input value to model neuron b is given by:

$$I(b) = \max_{a \in \mathcal{B}(\vec{x}_c)} (J_{ba} T''_{ba}), \tag{7}$$

Model graph blob $B(b_c)$ is activated according to the minimum potential value given by:

$$V(b_c) = -\sum_{b \in B(b_c)} I(b), \tag{8}$$

Let C_{b_c} denotes the set of feature correspondences between image neurons in $\mathcal{B}(\vec{x}_c)$ and model neurons in $B(b_c)$ that contributed to the activation values of $B(b_c)$ according to (7). Image neurons associated to feature correspondences used to calculate T''_{ba} for feature correspondences in C_{b_c} are shifted. The shift for one such image neurons say a^*, is given by:

$$\Delta\vec{x}_{a^*} = \kappa \mathcal{S}(P_{ba}, \vec{b^*}, \vec{a^*})(\vec{x}_{a^*} - \vec{x}^I_{a^*}), \tag{9}$$

where $\Delta\vec{x}_{a^*}$ denotes the shift in position of a^*. $b^* \in \mathcal{M}$ is the matching counterpart of a^*. $(\vec{x}_{a^*} - \vec{x}^I_{a^*})$ is the difference between the current position of a^* (denoted by $\vec{x}_{a^*}$) and the ideal position of a^* (denoted by $\vec{x}^I_{a^*}$). κ is the update rate, a value of 0.2 is used throughout our experiments. T_{ba} of all image neurons with direct intra-layer links to shifted image neurons are recomputed with equation (4).

If Λ_{b_c} represents the set of feature correspondences whose image neurons have direct intra-layer connections with image neurons in C_{b_c}, then J_{ba} is updated by:

$$J_{ba} = \frac{J_{ba} + \varepsilon J_{ba} T_{ba}}{\sum_{a'} (J_{ba'} + \varepsilon J_{ba'} T_{ba'})} \qquad \forall_{\{b,a\}} \in C_{b_c}, \Lambda_{b_c} \tag{10}$$

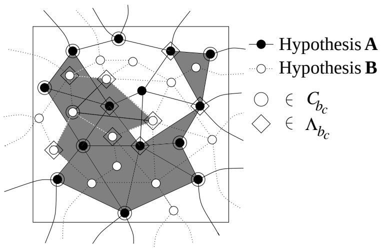


Figure 5: Competition for space between two overlapping hypotheses.

where the update rate ε is equated to 0.025 in current implementation. To deal with ambiguities caused by overlapping hypothesized matches as shown in Figure 5, we deliberately penalize some dynamic links within the activated blobs. More specifically, dynamic links connected to image neurons not found in C_{b_c} or Λ_{b_c} but lie within triangles (grayed areas) formed by connecting 3 image neurons with direct intra-layer links to each other (in C_{b_c} or Λ_{b_c}) are weakened by:

$$J_{ba} = J_{ba} - \gamma \exp\left(- \frac{(\vec{x}_a - \vec{\mu})^2}{5d^2} \right), \qquad (11)$$

where $\vec{x}_k$ is the position of image neuron k. $\vec{\mu}$ is the centroid of the overlap between image and model blobs. d is the maximum displacement from $\vec{\mu}$ among image neurons in C_{b_c} and Λ_{b_c}. A factor of $\gamma = 0.01$ is used in our experiments.

We used very large image blobs $\mathcal{B}(\vec{x}_c)$ at $\frac{1}{5}$ the test image size in the beginning of dynamic link matching, so that more probable hypothesized matches can be found rapidly, but gradually reduced the blob sizes with increasing iterations as dynamic links matured.

6. EXPERIMENTAL RESULTS

This section presents some exemplifying experiments which demonstrate the capabilities of the paradigm, namely invariance to perspective transformations and partial occlusion. Stage 1 is allowed to run for 800 iterations. Stage 2 terminates after 500 iterations or when $> 30\%$ of the model features are locked onto certain scene features for 70 consecutive iterations. Resultant matches are illustrated by superimposing model neurons onto the positions of image feature points which give the maximum activation values according to (7). Missing model features (possibly due to occlusion) with low input values ≤ 0.52 are not plotted.

Many test images are cluttered with irrelevant features which obscure straightforward detection of useful feature correspondences, especially in recognition of real objects which are exposed to physical noise caused by varying lighting conditions and shadows. In spite of these difficulties, the proposed system produces fairly accurate matches as shown in Figure 6. Most experiments generate the desired match within 300 iterations of stage 2.

7. CONCLUSION AND FUTURE WORK

A novel object-matching system based on dynamic link matching is reported. The proposed paradigm resolves several problems inherent in conventional neural-network-oriented matching techniques. Its merits include: invariance against perspective distortion, good tolerance against partial occlusion and against poor image segmentation. Reported experiments demonstrated the system's ability to produce remarkably accurate matches even when the scene is cluttered with irrelevant features.

Many man-made objects are either flat or have planar surfaces: thus, it is possible to recognize a 3D object by

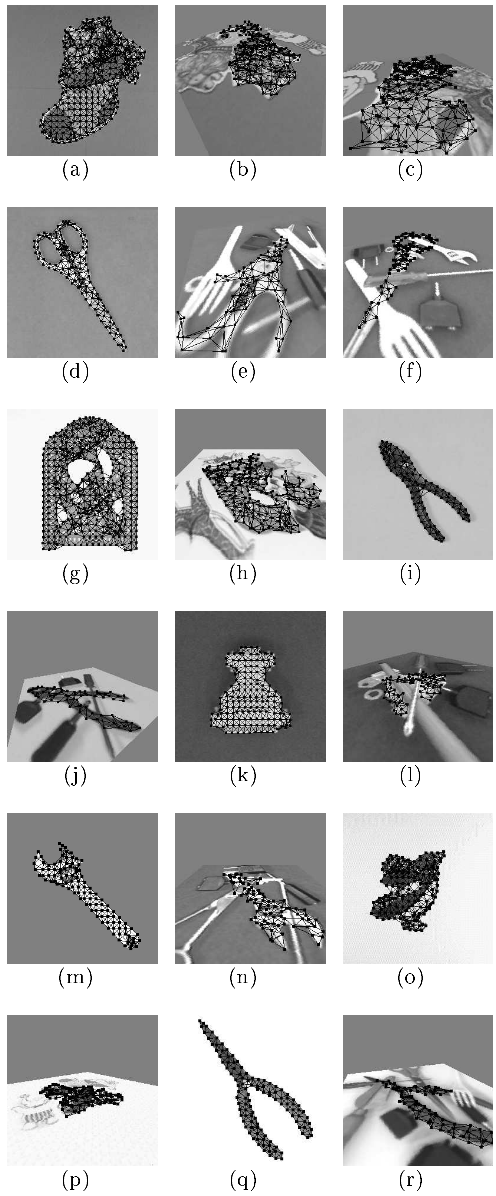


Figure 6: Experiment 1: (a) model grid, (b, c) resultant match. Experiment 2: (d) model grid and (e, f) resultant match. Experiment 3: (g) model grid and (h) resultant match. Experiment 4: (i) and (j). Experiment 5: (k) and (l). Experiment 6: (m) and (n). Experiment 7: (o) and (p). Experiment 8: (q) and (r).

matching one or several of its flat surfaces. The next phase of work will focus on extending this technique to matching 3D objects with planar surfaces.

8. REFERENCES

[1] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R.P. Würtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture", *IEEE Trans. Computers*, vol. 42, no. 3, pp. 300–310, March 1993.

[2] W.K. Konen, T. Maurer, and C. von der Malsburg, "A fast dynamic link matching algorithm for invariant pattern recognition", *Neural Networks*, vol. 7, no. 6/7, pp. 1019–1030, 1994.

[3] R.P. Würtz, J.C. Vorbrüggen, C. von der Malsburg, and J. Lange, "Recognition of human faces by a neuronal graph matching process", in *Applications of Neural Networks*, H. G. Schuster, Ed., pp. 181–200. VCH, Weinheim, Germany, 1992.

[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, London, 1990.

[5] T.M. Martinetz, S.G. Berkovich, and K.J. Schulten, "Neural-gas network for vector quantization and its application to time-series prediction", *IEEE Trans. Neural Networks*, vol. 4, no. 4, pp. 558–569, July 1993.