Recognition of Planar Objects in 3D Space

H. C. Sim and R. I. Damper

Department of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK
[hcs95r|rid]@ecs.soton.ac.uk

Abstract

A technique for recognizing planar objects in three-dimensional space is described. The object's domain is not restricted to purely two-dimensional items but includes fairly flat real objects such as a pair of scissors. To tackle realistic matching problems, the proposed method is invariant to changes within a prescribed range of perspective viewpoints, scaling, shift and reasonable degree of occlusion. In brief, objects are recognized by a modified dynamic link matching. Our experiments show that the system is very successful in recognizing deformed objects due to perspective distortion, even in rather cluttered scenes.

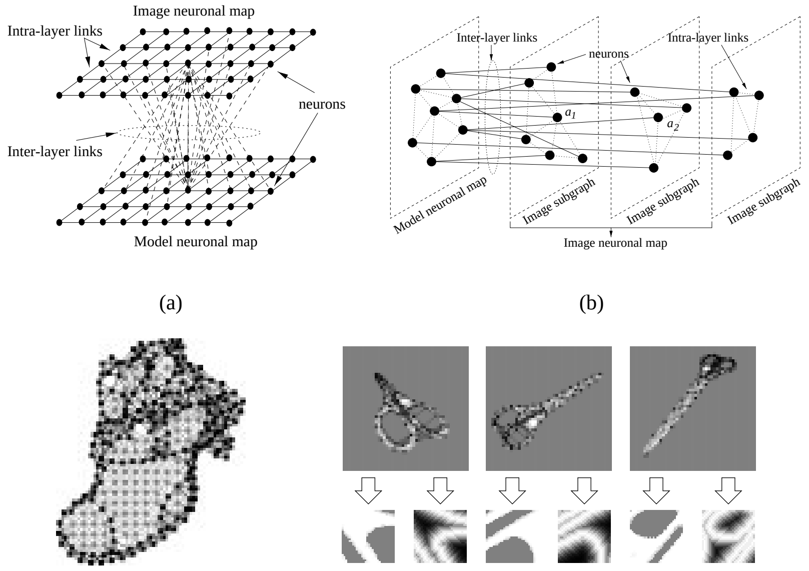
1 Introduction

The paper focuses on object matching – a high-level vision task which is an integral component of a machine-vision system. To date, most neural-network-oriented or conventional object matching systems are only invariant to affine transformation. In contrast, this paper takes a more generalized approach for planar object matching which is invariant to 3D perspective transformation and partial occlusion.

Dynamic link architecture (DLA), first proposed by von der Malsburg [2, 4] offers a new variant of neural processing which combines the merits of deformable models and self-organizing capability of neural networks. DLA involves two layers of rectangular neuronal maps labeled $\mathcal{I}$ (representing image graph) and $\mathcal{M}$ (representing model graph), as shown in Figure 1(a). Each neuron represents a local feature detector. There are intra-layer connections which encode the 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 links with the aid of appropriate intra- and inter-layer dynamics.

DLA has been applied very successfully in recognition of faces [2] and of partially occluded objects in cluttered scenes [4]. However, it has yet to be applied in recognizing objects in truly 3D scenarios. One crucial reason is that the appearances of local model features vary dramatically under perspective transformations resulting in an enormous

BMVC 1998 doi:10.5244/C.12.89



(a)

(b)

(c)

(d)

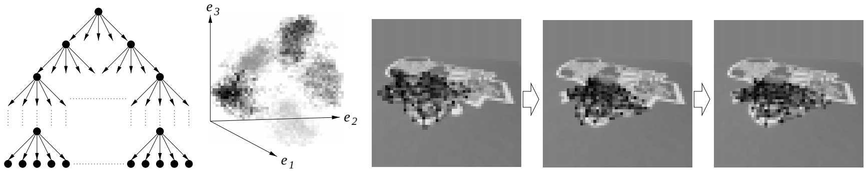
Figure 1: a) Classical DLA. b) Modified DLA. c) Model neuronal map. d) Varying views of model features.

search space. This work addresses the aforesaid problem using a self-organized hierarchical indexing system, so that a randomly-chosen test image feature only needs to be compared with a relatively small number of model features. Moreover, classical DLA is only tolerant to slight perspective distortions, e.g. small out-of-plane rotation. It cannot handle the large degrees of perspective deformation we try to accommodate here. Hence, several modifications to classical DLA are introduced to tackle this exceedingly difficult recognition task, which has being largely ignored by machine-vision researchers.

2 Feature Extraction and Model Representation

Figure 1(b) shows the architecture of the modified DLA. The proposed model neuronal map $\mathcal{M}$ is not a rectangular grid of neurons as illustrated in Figure 1(c). Unlike classical DLA (c.f. Figure 1(a)) where image and model neurons have all-to-all connections, an image neuron is connected to a single model neuron. However, several image neurons may terminate at the same model neuron. Furthermore, some image neurons connected to dissimilar model neurons may coincide in the test image, e.g. a_1 and a_2 in Figure 1(b). This unusual connectivity is chosen because adjacent model neurons may be mapped into the same point under 3D perspective transformations.

Gabor filter responses are selected as the set of local features represented by each neuron. The features are extracted by convolving the 2D image $I(\vec{x})$ with a set of 2D Gabor functions [2] given by:



(a) (b) (c)

Figure 2: a) Tree-like indexing system. b) 5 clusters formed at a root node. c) Progressive alignment of local features using simulated annealing and conjugate gradient minimization.

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

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 two different frequency levels, i.e. $w_\nu = \{0.27, 0.58\}$ and 8 different orientations, i.e. $\theta_\mu = \frac{\pi\mu}{8}$, $\mu \in \{0, \ldots, 7\}$. Radii R of the Gabor filters are 15 and 7 pixels for low and high frequencies respectively: the relative bandwidth is 2.0. Gabor wavelet representations are not useful for extract features on homogeneous surfaces. Hence, Gabor filters are applied to both gray level images and dilated edge maps as shown in Figure 1(d). The former is useful for highly-structured surface markings. The edge map is dilated by passing a Gaussian filter over the edge pixels. The system adopts a multiscale strategy to provide scale invariant matching between factors of 0.5 and 1.5.

3 Self-Organized Hierarchical Indexing System

A multi-view approach is used to provide 3D perspective invariant recognition. Gabor filter responses are extracted for model neurons when the object is seen from 140 evenly-spaced viewpoints (c.f. Figure 1(d)) between elevation angles of $20°$ and $90°$. A self-organized hierarchical indexing system (Figure 2) is created to group model features with almost identical Gabor responses. Hence, an unknown test image feature need only be compared with a few groups of closely-matching model features. For efficiency, only the low frequency filter responses are stored. Furthermore, model neurons are classified under *primary* or *secondary* features (c.f. Figure 1(c)). Primary features (shown as black dots) include sample points along object boundaries and other prominent contours. Secondary ones are evenly-spaced foreground sample points. Only primary features are stored: secondary features play the auxiliary role of supporting local matching of primary features.

Figure 2(a) shows the tree structure of our proposed indexing system. At each node, sampled Gabor filter responses are grouped as follows: 1) discrete Karhunen-Loéve expansion [1] is applied to extract effective features among Gabor responses; 2) highly correlated groups of these principal features are then generated based on unsupervised clustering; 3) the Fisher transform [1] is used to enhance class separability. The unsupervised clustering is basically an iterative k-means approach where each feature $\vec{x}$ is attracted to the nearest cluster according to its normalized Mahalanobis distance:

$$\mathcal{D}(\hat{G}, \vec{\mu}) = \frac{1}{2}\left(d \ln 2\pi + \ln|\Sigma| + (\hat{G} - \vec{\mu})^{\mathrm{T}} \Sigma^{-1} (\hat{G} - \vec{\mu})\right), \tag{2}$$

where $\hat{G}$ denotes the feature vector produced by Karhunen-Loéve expansion in step 1, $\vec{\mu}$ is the cluster centroid, d is the vector space dimensionality and Σ is the covariance matrix, whereby $|\Sigma|$ and Σ^{-1} denote its determinant and inverse respectively. The initial k centers for the k-means algorithm are derived using vector quantization based on "neural-gas" networks [3]. 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(b) shows 5 distinct clusters created at the root of a tree used to represent model neuronal map shown in Figure 5(a). 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 node. Two such indexing trees are created for Gabor responses of pixel intensity and dilated edge map (c.f. section 2).

4 Local Feature Matching and Alignment

At the leaves, edge gradient and principal orientation of local pixel intensity (determined by the axis of least second moment) are used to prune away model features unlikely to match the given image feature. Differences of both attributes for a pair of comparable model and test image features must be less than 20°. On average, a given image feature needs to match against less than 80 model features at the leaves. The similarity between an image feature $a \in \mathcal{I}$ and remaining model features $b \in \mathcal{M}$ are computed as:

$$\mathcal{S}(P_{ba}, \vec{b}, \vec{a}) = \frac{\vec{b} \cdot \vec{a}}{||\vec{b}||||\vec{a}||} \min\left(\frac{||\vec{b}||}{||\vec{a}||}, \frac{||\vec{a}||}{||\vec{b}||}\right), \tag{3}$$

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, and 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 ≥ 0.75 in equation (3) 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 responses with equation (3). 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. 7 perspective parameters must be optimized: $P_{ba} =$ elevation angle (θ_{ev}), azimuth angle (θ_{az}), horizontal shift (d_x), vertical shift (d_y), zoom factor s, and view reference point co-ordinates (v_x, v_y). The cost function is formulated as:

$$\mathcal{C}(P_{ba}) = \frac{1}{N} \sum \left(1 - e^{-\alpha D(\vec{x})} \left| \cos(\triangle(\vec{x})) \right| \right), \tag{4}$$

where $D(\vec{x})$ is the distance from 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 transformation is used here mainly because of insensitivity to small changes in P_{ba}. A two-stage optimization process is used: 1) simulated annealing (SA) with downhill simplex minimization is used to align feature correspondences quickly; 2) transformation parameters produced by SA are then refined by conjugate gradient minimization. The 8 vertices on the simplex are initialized to implied perspective transformations of the 8 most likely matching feature correspondences. Hence, competition among hypothesized matches occurs at this level. Partial derivatives of the translation parameters (d_x, d_y) are approximated by finite differences, whereas partial derivatives of the remaining parameters are determined by the chain rule using partial derivatives of d_x and d_y with respect to the other transformation parameters.

Only edges surrounding L levels of projected model neurons in the test image are used for feature alignment. Figure 2(c) shows this process for $L = 5$ levels of model neurons. Both primary and secondary features are taken into account.

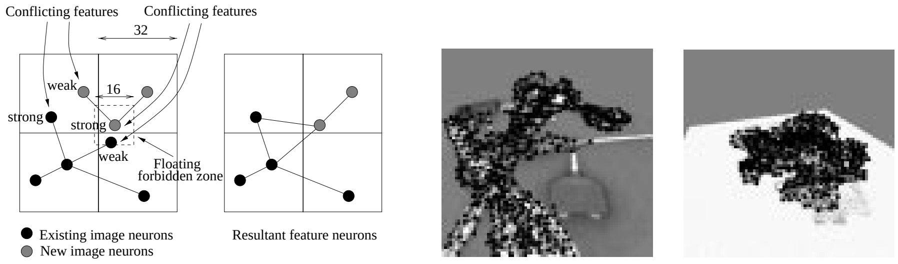
5 Dynamic Link Matching

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 two: first, the neuronal map $\mathcal{I}$ is created; in the second stage, our modified dynamic link matching process is applied to find the desired match.

5.1 Stage One: Creation of Image Neuronal Map

Edge pixels are activated in a square window denoted by $\mathcal{W}(\vec{x})$ centered at a randomly selected point $\vec{x}$ in the test image. Matching counterparts are determined for activated edge points using the indexing tree (which contains only primary features associated with model edge points). Simulated annealing and conjugate gradient minimization (c.f. 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. Hence, local subgraphs of feature correspondences are created for both primary and secondary features after the alignment process. If $g(a^*) \in \mathcal{M}$ represents the model counterpart of image neuron $a^* \in \mathcal{I}$, then local similarity measures $T_{ba} = \{T'_{ba}, T''_{ba}\}$ are calculated for newly-created feature correspondence $\{b \in \mathcal{M}, a \in \mathcal{I}\}$ as follows:

$$\begin{aligned} T'_{ba} &= \mathcal{S}(P_{ba}, \vec{b}, \vec{a}) \\ T''_{ba} &= \lambda T'_{ba} + (1-\lambda) h_{ba}, \end{aligned} \tag{5}$$



(a) (b)

Figure 3: a) Forbidden zones used to resolve conflicting feature correspondences. b) Image neuronal maps generated in stage one.

$$h_{ba} = \frac{1}{2|\Lambda_b|} \sum_{b^* \in \Lambda_b} \max_{\substack{a^* \in \Lambda_a \\ g(a^*) = b^*}} \left[\mathcal{S}(P_{ba}, \vec{b}^*, \vec{a}^*) + \vec{\mathcal{X}}_{aa^*} \right] \qquad (6)$$

where $\mathcal{S}(\cdot)$ is given in equation (3), h_{ba} measures feature conformity within the neighborhood of b and a, Λ_k is the set of neurons directly connected to neuron k via intra-layer links, $|\Lambda_k|$ 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^*, and λ (= 0.7 here) dictates the relative significance of T'_{ba} and h_{ba}. T'_{ba} applies to the outermost level L of model neurons used in the alignment process, whereas T''_{ba} applies to model neurons between levels 0 and $L - 1$.

Numerous local subgraphs of feature correspondences are created with random activation of $\mathcal{W}(\vec{x})$. A growth control mechanism based on competition and co-operation among local match hypotheses (or subgraphs of image neurons) prevents proliferation of spurious feature correspondences while promoting the merger of comparable image subgraphs. Forbidden zones are defined whereby no two image neurons are allowed to converge at the same model neuron. There are two types of forbidden zone: fixed and floating. A 256×256 test image is divided into 64 non-overlapping 32×32 fixed forbidden zones as in Figure 3. Floating forbidden zones are 16×16 windows surrounding newly-created image neurons where the uniqueness constraint also applies (one such is shown in Figure 3). When ≥ 2 image neurons converge at the same model neuron in these forbidden zones, only the one with highest Gabor response similarity given by equation (3) is retained. Intra-layer links connected to the expunged image neurons are reconnected to the selected image neurons (e.g. Figure 3). T_{ba} is recalculated for affected feature correspondences which have modified intra-layer connections using (5). The number of image neurons will not exceed $64 \times |\mathcal{M}|$, where $|\mathcal{M}|$ is the number of model neurons.

Here, activation windows $\mathcal{W}(\vec{x})$ are of size 32×32 and $L = 3$ levels of model neurons are used in the local alignment. Smaller $\mathcal{W}(\vec{x})$ allows discovery of more dissimilar match hypotheses by minimizing the likelihood of selecting the same winning feature correspondences repeatedly. Larger L will increase interaction between hypotheses so as to provide more evidence (or feature correspondences) to warrant a plausible match eventually.

The resultant image neuronal map $\mathcal{I}$ is a complex network of subgraphs created by many independent hypothesized matches as shown in Figure 3(b). Most of these image subgraphs are connected at some points because of the interactions among hypothesized matches described above. However, we do not rule out the possibilities of disjoint image neuronal maps $\mathcal{I}$. Moreover, some of these disjoint maps may even be overlapping.

5.2 Stage Two: Modified DLA

Inter-layer dynamic links $J_{ba} \in [0, 1]$ are established for pairs of potentially matching model neuron b and image neuron a according to:

$$J_{ba} = \frac{T''_{ba}}{\sum_{a'} T''_{ba'}}, \tag{7}$$

where a' are image neurons connected to b. An image blob denoted by $\mathcal{B}(\vec{x}_c)$ is randomly activated as part of the system dynamics. Unlike classical DLA, image blobs used here are simply square windows centered at randomly selected test image point $\vec{x}_c$. We cannot use the same image blob definition as in conventional DLA for two reasons: 1) the image neuronal map is not a rectangular grid; 2) there may be overlaps among image subgraphs as shown in Figure 4.

Effective input value to model neuron b is given by:

$$I\left(b\right) = \max_{a \in \mathcal{B}(\vec{x}_c)} (J_{ba} T''_{ba}), \tag{8}$$

Only $I\left(b\right) > 0.52$ are considered as probable matches. Model graph blob $B\left(b_c\right)$ is activated according to the minimum potential value given by:

$$V\left(b_c\right) = - \sum_{b \in B(b_c)} I(b), \tag{9}$$

Let C_{b_c} be 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 equation (8). Image neurons associated with feature correspondences used to calculate T''_{ba} for feature correspondences in C_{b_c} are shifted. The shift for one such image neuron, say a^*, is given by:

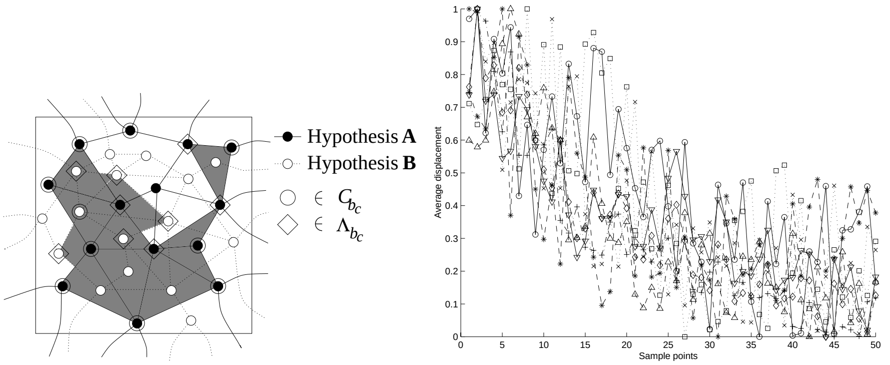
$$\vec{x}_{a^*} = \vec{x}_{a^*} + \kappa \mathcal{S}(P_{ba}, \vec{b^*}, \vec{a^*})(\vec{x}_{a^*} - \vec{x}^I_{a^*}), \tag{10}$$

where $\vec{x}_{a^*}$ denotes the position of image neuron a^* and b^* is the matching counterpart of a^*, $(\vec{x}_{a^*} - \vec{x}^I_{a^*})$ is the displacement between the current position of a^* and the ideal position of a^* (denoted by $\vec{x}^I_{a^*}$) and κ is the update rate. A value of 0.2 is used throughout our experiments. T_{ba} of all adjacent image neurons with intra-layer links to shifted image neurons is recalculated with equation (5).

Let Λ_{b_c} denote the set of feature correspondences whose image neurons have intra-layer connections with image neurons in C_{b_c}. Then J_{ba} is updated by:

$$J_{ba} = \frac{J_{ba} + \varepsilon J_{ba} T_{ba}}{\sum_{a'} \left(J_{ba'} + \varepsilon J_{ba'} T_{ba'}\right)} \qquad \forall_{\{b,a\}} \in C_{b_c}, \Lambda_{b_c} \tag{11}$$

where the update rate ε is equated to 0.025 here. Disjoint image neuron subgraphs associated with different hypotheses (i.e. dissimilar perspective transformations) may occupy the same spatial region in the test image as shown in Figure 4. To resolve ambiguities caused by overlapping image neuron subgraphs, we deliberately weaken some dynamic links within the activated blobs. Links connected to image neurons not found in C_{b_c} or



(a)

(b)

Figure 4: a) Competition between two overlapping hypotheses. b) Average displacements of model neurons per 10 iterations during dynamic link matching

Λ_{b_c} but encompassed by triangles (grayed areas) formed by connecting 3 image neurons with intra-layer links to each other (in C_{b_c} or Λ_{b_c}) are weakened by:

$$J_{ba} = J_{ba} - \gamma \exp\left(- \frac{(\vec{x}_a - \bar{x})^2}{5x_m^2} \right), \tag{12}$$

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

Initially, image blobs $\mathcal{B}(\vec{x}_c)$ at $\frac{1}{5}$ the test image size are used, so that more probable hypothesized matches can be found rapidly. Blobs gradually shrink with increasing iterations as dynamic links mature.

6 Experimental Results

This section describes some experiments which demonstrate the capabilities of the described paradigm, namely invariance to perspective transformations and robustness to noise 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 test image sample points 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. Missing model features (possibly resulting from partial occlusion) with low input values (≤ 0.52) are not shown.

The range of objects used in testing varies from 2D planar items to flat real objects. Many test images are cluttered with irrelevant features which obscure straightforward detection of useful feature In spite of this, the proposed system produces fairly accurate matches as shown in Figure 5.

The proposed system should not produce large networks of connected image neurons when a given target is *not* present in the test image. Figures 6(a) and (b) show the resultant matches when the system tried to locate the model in Figure 5(a) in the given test images.

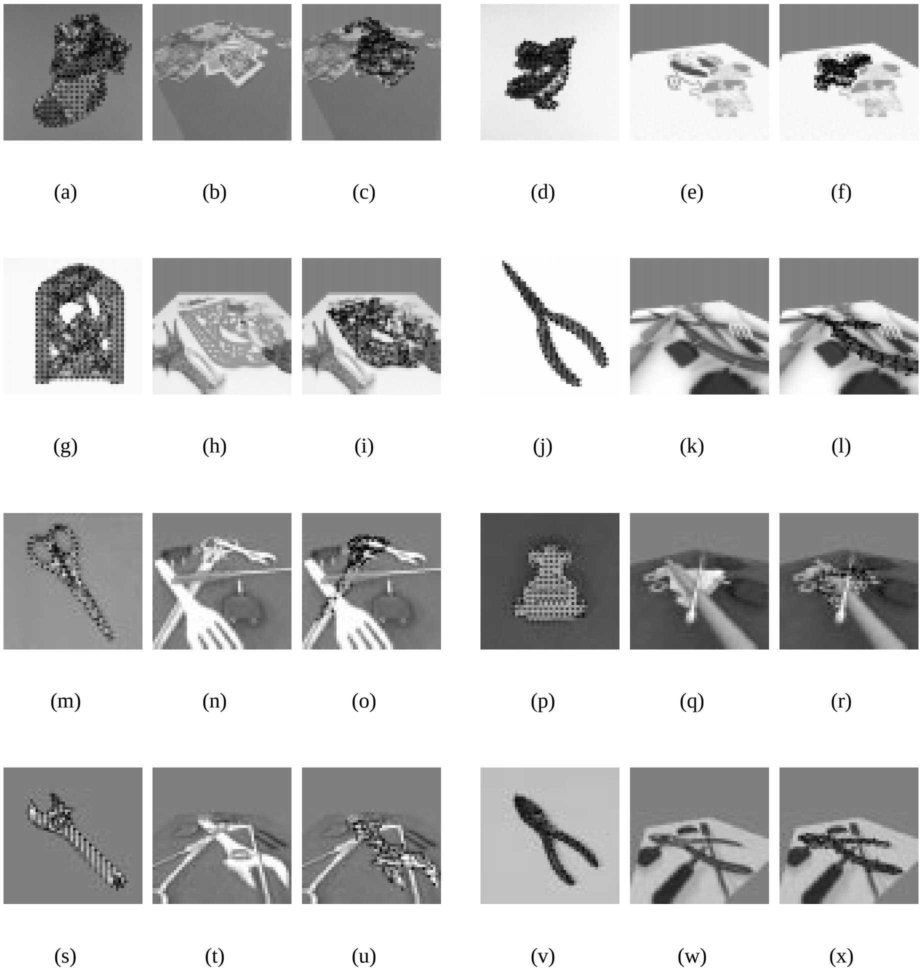


Figure 5: (a, d, g, j, m, p, s, v): model neuronal maps. (b, e, h, k, n, q, t, w): test images. (c, f, i, l, o, r, u, x): resultant matches.

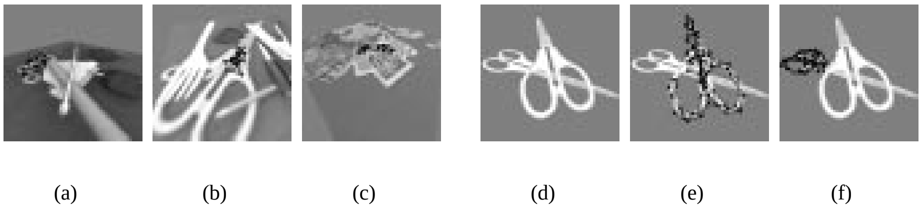


Figure 6: (a, b, c) Small model graphs generated by proposed system when the target is not present in the test images. (d, e, f) Applying the proposed system to recognize multiple occurrences of a target.

Likewise, a poor match is obtained when the model in Figure 5(m) is applied to the test image shown in Figure 6(c). We have also applied the algorithm to images containing more than one instance of a given target, e.g. Figure 6(d). Figures 6(e) and (f) show the largest and second largest image subgraphs produced respectively, demonstrating the potential to recognize multiple occurrences of identical objects in test images.

Figure 4(b) shows the average displacement of model neurons due to shifting or reassignment of their most likely matching counterparts in the test image during dynamic link matching. Average displacements plotted in the graphs are normalized to between the smallest and largest displacements measured for each experiment. 50 measurements are taken at 10 iterations apart. Most of the experiments show rather erratic changes in average displacements throughout matching process. However, many of them show gradual decrease in average displacement (or downward trend) with increasing iterations.

7 Concluding Remarks and Future Work

A neural technique for recognizing flat objects is described. The merits of this approach over most conventional neural-network based recognition systems include: invariance against perspective distortion, good tolerance against partial occlusion and poor image segmentation. The next phase of work focuses on extending this approach to recognize 3D artificial objects by detecting one or more of their planar surfaces. We also intend to adapt the current implementation to match multiple occurrences of identical objects in test images.

References

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

[2] 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 Transactions on Computers*, 42(3):300–310, March 1993.

[3] T.M. Martinetz, S.G. Berkovich, and K.J. Schulten. Neural-gas network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569, July 1993.

[4] L. Wiskott and C. von der Malsburg. A neural system for the recognition of partially occluded objects in cluttered scenes : A pilot study. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(4):935–948, 1993.