

Computation in the Higher Visual Cortices: Map-Seeking Circuit Theory and Application to Machine Vision

David Arathorn
Center for Computational Biology
Montana State University
and
General Intelligence Corporation
PO Box 7380, Bozeman, MT, 59771
dwa@giclab.com

Abstract

Map-Seeking Circuit theory is a biologically based computational theory of vision applicable to difficult machine vision problems such as recognition of 3D objects in arbitrary poses amid distractors and clutter, as well as to non-recognition problems such as terrain interpretation. It provides a general computational mechanism for tractable discovery of correspondences in massive transformation spaces by exploiting an ordering property of superpositions. The latter allows a set of transformations of an input image to be formed into a sequence of superpositions which are then “culled” to a composition of single mappings by a competitive process which matches each superposition against a superposition of inverse transformations of memory patterns. The architecture that performs this is based on a number of neuroanatomical features of the visual cortices, including reciprocal dataflows and inverse mappings.

1. Introduction

The mechanism described here evolved from an effort to “reverse engineer” the visual cortices to create a viable machine vision mechanism. Many of the techniques used in conventional computational vision are precluded either by the limitations of plausible neuronal circuitry or by psychophysical characteristics of biological vision. For example, it can be reasonably argued that the repertoire of computations that can be implemented by realistic neurons is limited to combination (summing of signals from dendritic branches, linear over a limited range) and analog gating (pseudo-multiplication locally in thin branches of dendrites)[1], thresholding and clipping, competition (via lateral inhibition) and mapping (via synaptic interconnectivity patterns). On the other hand, neuroanatomy offers a rich hint of a solution in the vast neuronal resources allocated to creating reciprocal top-

down and bottom-up pathways. More specifically, this reciprocal pathway architecture appears to be organized with reciprocal, co-centered fan outs in the opposing directions [2].

Visual psychophysics provides such a vast array of hints as to how the mammalian visual system is organized that it is difficult to organize them. However, individual psychophysical behaviors preclude some of the favored computational approaches of machine vision. For example, the variety of kinetic depth effects which allow us to distinguish 3D surfaces defined by moving sets of indistinguishable dots makes any solution by determining pairwise correspondences virtually intractable. Similarly, our ability to recognize distant objects which the fovea can represent with less than a dozen cycles of resolution is a psychophysical behavior which precludes the invariant feature approaches often used in machine vision.

The hope of biomimetic vision is that by discovering and applying the principles which drive biological vision we may at least approach if not exceed the capabilities of mammalian vision. Thus neuroanatomy, neurophysiology and visual psychophysics must be allowed to provide both the constraints and treasure map for the search.

The general mechanism described here was evolved by this strategy. It has come to be called a *map-seeking circuit* because its mathematical expression has an isomorphic implementation in quite realistic neuronal circuitry [3]. The most striking features of the architecture of both, seen in Figure 1, are the reciprocal forward and backward pathways – a signature of the biological cortices – with their forward and inverse mapping sets. In mathematical form, as seen in equations 6-9 below, it provides a parsimonious, computationally explicit theoretical mechanism for the accumulating body of neurobiological evidence that top-down expectations drive vision – a conclusion also imposed by mathematical necessity in what is otherwise an ill-posed problem. The practical effectiveness of the same equations will also be demonstrated.

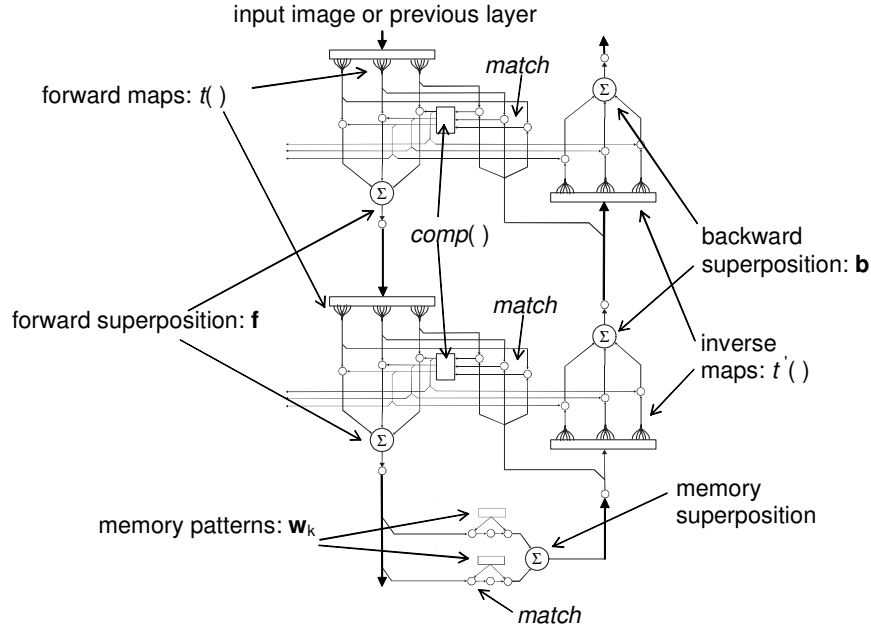


Figure 1. Data flow in map-seeking circuit

2. The problem

The abstract problem solved by the map-seeking circuit is the discovery of a composition of transformations between an input pattern and a stored pattern (or between two input patterns, as in the case of stereovision). In general the transformations express the generating process of the problem. Define correspondence c between vectors \mathbf{r} and \mathbf{w} through a composition of L transformations $t_{j_1}^1, t_{j_2}^2, \dots, t_{j_L}^L$ where $t_{j_l}^l \in t_1^l, t_2^l, \dots, t_m^l$

$$c(\mathbf{j}) = \left\langle \begin{matrix} L \\ \circ \\ i=1 \end{matrix} t_{j_i}^i(\mathbf{r}), \mathbf{w} \right\rangle \quad \text{eq. 1}$$

where the composition operator is defined

$$\begin{matrix} L \\ \circ \\ i=0,1 \end{matrix} t_{j_i}^i(\mathbf{r}) = \begin{cases} l=1 \dots L & t_{j_l}^l \circ t_{j_{l-1}}^{l-1} \dots \circ t_{j_1}^1(\mathbf{r}) \\ l=0 & \mathbf{r} \end{cases}$$

Let \mathbf{C} be an L dimensional matrix of values of $c(\mathbf{j})$ whose dimensions are $n_1 \dots n_L$. The problem, then is to find

$$\mathbf{x} = \arg \max c(\mathbf{j}) \quad \text{eq. 2}$$

The indices \mathbf{x} specify the sequence of transformations that best correspondence between vectors \mathbf{r} and \mathbf{w} . The problem is that \mathbf{C} is too large a space to be searched by conventional means. Instead, map-seeking circuits search a superposition space Q defined

$$Q: \mathbb{R}^{\sum_{i=1}^m} \rightarrow \mathbb{R}^1$$

$$Q(\mathbf{G}) = \left\langle \begin{matrix} m-1 \\ \circ \\ i=1 \end{matrix} \left(\sum_i g_i^i \cdot t_i^i \right) (\mathbf{r}), \begin{matrix} m+1 \\ \circ \\ i=l=0, L \end{matrix} \left(\sum_i g_i^i \cdot t_i^i \right) (\mathbf{w}) \right\rangle \quad \text{eq. 3}$$

where $\mathbf{G} = [g_{x_m}^m]$ $m = 1 \dots L, x_m = 1 \dots n_m$ n_m is number of t in layer m , $g_{x_m}^m \in [0, 1]$, t_i^l is adjoint of t_i^l .

$Q(\mathbf{G})$ is the hypersurface defining the value of the inner product of forward and backward superpositions for all values of g . In Q space, the solution lies along a single axis in each layer. *Superposition culling* uses the components of $\text{grad } Q$ to compute a path in steps Δg to the axis in each layer l which corresponds to the best fitting transformation t_{x_l} , where x_l is the l^{th} element of \mathbf{x} in eq. 2.

$$\frac{\partial Q(\mathbf{G})}{g_j^m} = \left\langle t_j^m \begin{matrix} m-1 \\ \circ \\ i=0,1 \end{matrix} \left(\sum_i g_i^i \cdot t_i^i \right) (\mathbf{r}), \begin{matrix} m+1 \\ \circ \\ i=l=0, L \end{matrix} \left(\sum_i g_i^i \cdot t_i^i \right) (\mathbf{w}) \right\rangle \quad \text{eq. 4}$$

$$\Delta \mathbf{g}^m = f \left(\frac{\partial Q(\mathbf{G})}{\partial g_1^m}, \dots, \frac{\partial Q(\mathbf{G})}{\partial g_{n_m}^m} \right) \quad \text{eq. 5}$$

The function f preserves the maximal component and reduces the others: in neuronal terms, lateral inhibition. This reformulation of the problem into the superposition space Q permits a search with resources proportional to the sum of sizes of the dimensions of \mathbf{C} instead of their product.

The price for moving the problem into superposition space is that *collusions* of components of the superpositions can result in better matches for incorrect mappings than for the mappings of the correct solution. The ordering property of superpositions [4] gives a probabilistic description of the occurrence of collusion for pattern vectors which satisfy the distribution properties of decorrelating encodings, for which there is reasonable

neurobiological evidence [5]. The formal statement of the superposition ordering property has two cases:

a) If a superposition $\mathbf{s} = \sum_{i=1}^n \mathbf{v}_i$ is formed from a set of sparse vectors $\mathbf{v}_i \in \mathbf{V}$, then for a vector \mathbf{v}_k which is not part of the set from which superposition is formed, $\mathbf{v}_k \notin \mathbf{V}$, the following relationship expresses the ordering property of superpositions:

$$P_{correct} > P_{incorrect}$$

$$\text{where } P_{correct} = P(\mathbf{v}_i \bullet \mathbf{s} > \mathbf{v}_k \bullet \mathbf{s}),$$

$$P_{incorrect} = P(\mathbf{v}_i \bullet \mathbf{s} \leq \mathbf{v}_k \bullet \mathbf{s})$$

and as $n \rightarrow 1$

$$P_{correct} \rightarrow 1$$

b) If three superpositions

$$\mathbf{r} = \sum_{i=1}^n \mathbf{u}_i, \mathbf{s} = \sum_{j=1}^m \mathbf{v}_j \quad \text{and} \quad \mathbf{s}' = \sum_{k=1}^m \mathbf{v}_k$$

are formed from three sets of sparse vectors $\mathbf{u}_i \in \mathbf{R}$, $\mathbf{v}_j \in \mathbf{S}$ and $\mathbf{v}_k \in \mathbf{S}'$ where $\mathbf{R} \cap \mathbf{S} = \emptyset$ and $\mathbf{R} \cap \mathbf{S}' = \mathbf{v}_q$ then the following relationship expresses the superposition ordering property:

$$P_{correct} > P_{incorrect}$$

$$\text{where } P_{correct} = P(\mathbf{r} \bullet \mathbf{s}' > \mathbf{r} \bullet \mathbf{s}),$$

$$P_{incorrect} = P(\mathbf{r} \bullet \mathbf{s}' \leq \mathbf{r} \bullet \mathbf{s})$$

and as $n, m \rightarrow 1$

$$P_{correct} \rightarrow 1$$

The proof [6] of these makes use of the assumption of pdf or pmf equality

$$f_i(q) = f_k(q) \text{ where}$$

$$f_i(q) = P(\mathbf{v}_i \bullet (\mathbf{s} - \mathbf{v}_i) = q), \quad f_k(q) = P(\mathbf{v}_k \bullet (\mathbf{s} - \mathbf{v}_i) = q)$$

For a given sparsity of image encoding the probability of the occurrence of collusion decreases with the decrease in numbers of contributing components in the superposition(s). Thus the strategy for exploiting the superposition space reliably is to allow convergence to quickly prune the number of contributors to the superpositions before committing to a solution. For problem spaces which make collusion less likely the convergence can be allowed to proceed more quickly. At the limit, in problem spaces which preclude collusion, the first iteration gives the solution.

The role of the ordering property in convergence to a correct solution can be understood in how it dictates the shape of the surface Q , specifically the relationship of the components of $\text{grad } Q$ in eq. 5. That is, the largest component of $\text{grad } Q$ in each layer will correspond with the correct mapping with increasing probability where more of the elements of \mathbf{G} are near zero.

The decomposition of the aggregate transformation into subtransformations proves to be more than a

mathematical convenience. The specific characteristics of the subtransformations turn out to carry cognitively critical information. In vision mapping classes specify pose, distance, and location in the visual field or yield the parameters for surface orientation for shape-from-view-displacement computation.

3. The map-seeking solution

A map-seeking circuit is composed of several transformation or mapping layers between the input at one end and a memory layer at the other. Other than the pattern of individual connections which implement the set of mappings in each layer, the layers themselves are more or less identical. The compositional structure is evident in the simplicity of the equations (eqs. 6-9 below) which define a circuit of any dimension.

In a multi-layer circuit of L layers plus memory with n_l mappings in layer l the mapping coefficients g are updated by the recurrence

$$g_i^m := \text{comp}(g_i^m, t_i^m(\mathbf{f}^{m-1}) \bullet \mathbf{b}^{m+1}) \text{ for } m=1 \dots L, i=1 \dots n_l \text{ eq. 6}$$

where match operator $\mathbf{u} \bullet \mathbf{v} = q$, q is a scalar measure of goodness-of-match between \mathbf{u} and \mathbf{v} , and \mathbf{f} and \mathbf{b} are defined below. When \bullet is a dot product the second argument of comp is $\partial Q/g$ in eq. 4. The function comp is a realization of lateral inhibition function f in eq. 5.

$$\text{comp}(g_i, q_i) = \max\left(0, g_i - k_1 \cdot \left(1 - \frac{q_i}{h}\right)^{k_2}\right) \text{ eq. 7}$$

$$\text{where } h = \begin{cases} \max \mathbf{q} & \text{if } \max \mathbf{q} > t \\ t & \text{if } \max \mathbf{q} \leq t \end{cases}$$

The forward path signal for layer m is computed

$$\mathbf{f}^m = \sum_{j=1}^{n_l} g_j^l \cdot t_j^l(\mathbf{f}^{m-1}) \quad \text{for } m=1 \dots L \text{ eq. 8}$$

The backward path signal for layer m is computed

$$\mathbf{b}^m = \begin{cases} \sum_{j=1}^{n_l} g_j^l \cdot t_j^l(\mathbf{b}^{m+1}) & \text{for } m=1 \dots L \\ \sum_k z(\mathbf{w}_k \bullet \mathbf{f}^L) \cdot \mathbf{w}_k & \text{for } m=L+1 \end{cases} \text{ eq. 9}$$

In above, \mathbf{f}^0 is the input signal, t_j^l, t_j^l are the j^{th} forward and backward mappings for the l^{th} layer, \mathbf{w}_k is the k^{th} memory pattern, $z(\cdot)$ is a non-linearity applied to the response of each memory. \mathbf{g}^l is the set of mapping coefficients g_j^l for the l^{th} layer each of which is associated with mapping t_j^l and is modified over time by the expression that is the second argument of the competition function. $\text{comp}(\cdot)$ is any function (one is illustrated) which leaves the maximum g in \mathbf{g}^l unchanged and moves the other values of g toward zero in proportion to their difference from the maximum element. Constants k_1 and k_2 control the speed of convergence.

The mapping gain coefficient g , is a relative measure of the probability that the singleton pattern corresponds to a contributor to the superposition, or that the two superpositions both contain a corresponding pattern. The competition process guarantees that each layer will converge to a single mapping with a non-zero g coefficient if the input pattern can be matched to one of the memory patterns, or that it will converge to all zero coefficients for all mappings in all layers if no acceptable match (i.e. sub-threshold) can be established. This process of superposition culling converges in time (or steps) largely independent of the number of initially active mappings.

It is the ordering property of superpositions, discussed above, that causes the best set of mappings ultimately to be selected with high probability at the end of convergence. This probability increases as the convergence proceeds since it is inversely related to the number of contributors to the superposition. Since there is a non-zero probability of selection of a memory pattern or transformation which provides a poor match when in fact a good match is available, the failure mode of the map-seeking circuit is to converge to a false negative. In practice this is rarely seen.

Recognizing 2D projections of 3D objects

The space of transformations with which a 3D object may project onto the retina, even when limited by discretization and limitation of range, in practice exceeds 10^{10} . By usual methods this space is intractably large to search exhaustively. However, the use of superpositions and decomposition, as proposed in map-seeking circuit theory, allows this space to be searched in biologically (and technologically) realistic time. An example of a recognition problem in this domain is seen in Figure 2: given a 3D model of a particular model of tank, distinguish the correct target undeterred by occlusion, noise and distractors.

The map-seeking circuit used in the demonstration in Figs. 2, 3 has four layers of transformational mappings: translation (120×120 pixels by steps of 1 pixel), rotation in plane (-15° to $+15^\circ$ by 1°), scaling (0.4 to 1.4 by steps of 0.025) and 3D projection (azimuth -90° to $+90^\circ$ by 5° , elevation 0° to 60° by 5°). A zero-th, initial, layer is used to process the gray scale image into an oriented-edge representation which is used in the 2D domain layers.

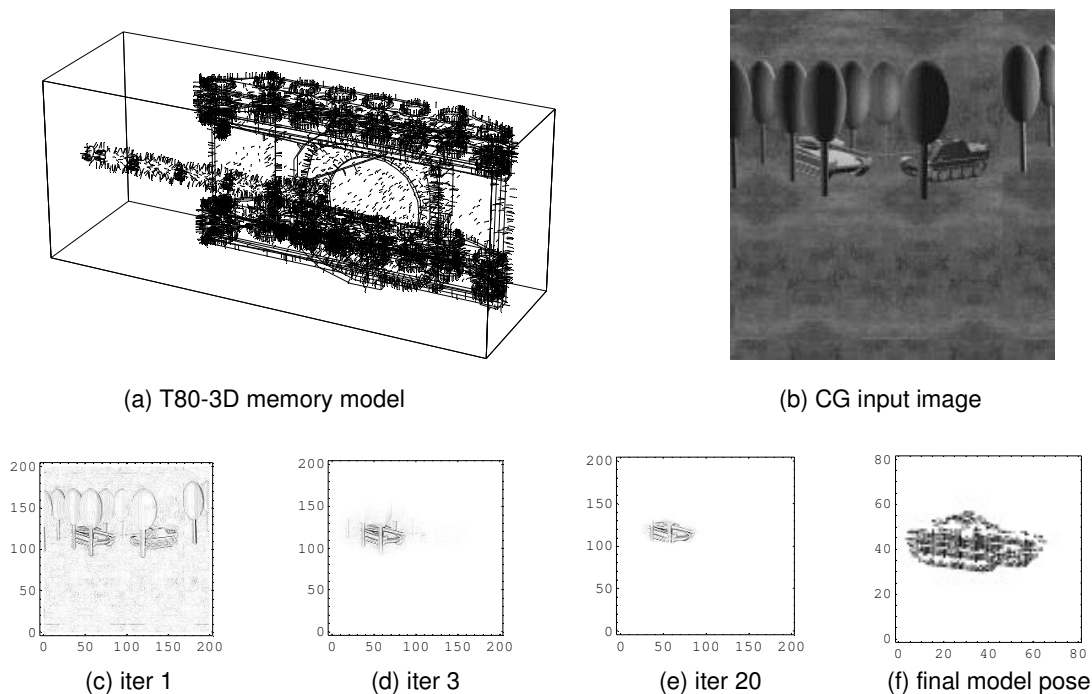


Figure 2. Occluded target (T-80) with similar distractor, computer generated scene. (a) T80-3D memory model encoded as normals and edges; (b) CG input image: occluded target and distractors; (c-e) isolation of target in layer 0, iterations 1, 3, 20; (f) determination of pose in final iteration, layer 4 backward.

This zero-th layer also provides the capability to operate in very low resolution and noisy conditions, as will be discussed below.

Figure 2c-e illustrate the evolution of the forward signal from layer 0 during normal convergence of the circuit to the correct sequence of transformations in layers 1-4 respectively, which determine the location, scale, rotation-in-viewing-plane and (nearly) correct 3D pose of the target. Figure 2f is the final projection of the model.

Recognition in real operating conditions

In real environments, biological as well as machine vision is called upon to identify objects at distances or in conditions which limit the resolution to 10 or fewer cycles on the long axis of the object. At such low resolutions there are no alignable features other than the shape of the object itself, and even its own boundaries are sufficiently blurred as to prevent generating reliable edges in a feed-forward manner. However, the 3D model in memory can be used to hypothesize top-down, in biological parlance, possible locations of edges in blurred image. The inverse-mappings of the 3D model to the possible projections, scalings, rotations and translations creates a set of edge

hypotheses on the backward path out of layer 1 into layer 0. In layer 0 these hypotheses are used to gate the input image. As convergence proceeds, the edge hypotheses are reduced to a single edge hypothesis that best fits the grayscale input image.

Figure 3 shows one of a series of tests of the same circuit used in the demonstration in Figure 2 applied to deliberately blurred images from the Fort Carson Imagery Data Set. The 3D memory model used in these tests is a simplified representation of an M-60 tank without a barrel. The circuit has no difficulty distinguishing the location and orientation of the tank, despite distractors and background clutter.

Performance

On computer generated imagery, such as seen in Figure 2, the map-seeking circuit always located the correct target unless deliberately induced to fail by severely reducing target contrast relative to the contrast of a similar non-target vehicle or by moving it to an extreme peripheral location relative to distractor. In all cases the circuit was able to determine orientation to about $\pm 15^\circ$. Using the Fort Carson visual spectrum input imagery, both blurred

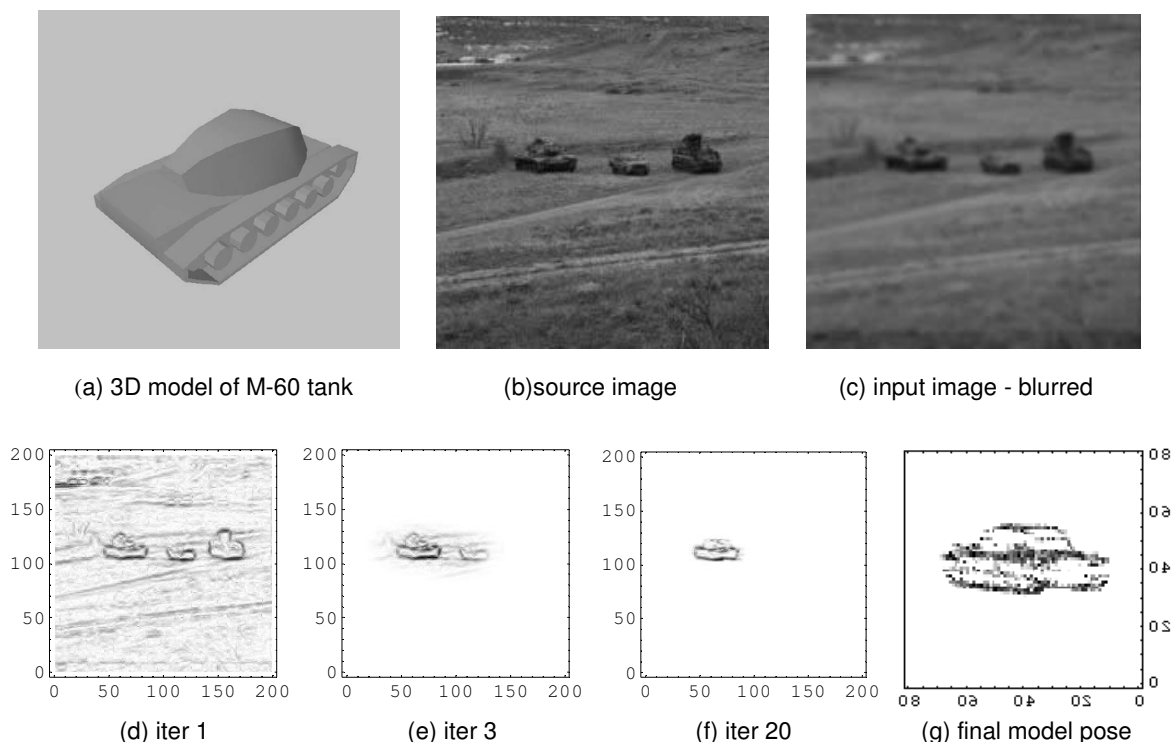


Figure 3. Target (M60) with distractor vehicles, Fort Carson scene. (a) M60 3D memory model; (b) source image; (c) Gaussian blurred input image; (d-f) isolation of target in layer 0, iterations 1, 3, 20; (g) pose determination in final iteration, layer 4 backward. Note model pose is presented left-right mirrored to reflect mirroring determined in layer 3. M-60 model courtesy Colorado State University.

and full resolution, the map-seeking circuit always located the target and in about 60% of cases correctly determined orientation to about $\pm 15^\circ$. Most of the incorrect orientations were “symmetries” of the correct orientation, i.e. projections with very similar silhouette to the correct projection. On Fort Carson IR imagery, some incorrect identifications of vehicles occurred with low contrast imagery. When contrast was normalized and IR hotspots removed, performance was similar to performance on visual spectrum imagery.

4. Conclusion

Object recognition from 3D memory models proves to be a natural application for map-seeking circuits, particularly in operating conditions that have proved extremely difficult for other methods, as exemplified by the low-resolution or otherwise degraded imagery illustrated in Figure 3. This is a compelling indication that drawing inspiration from plausible biological visual circuitry is a path of research for machine vision that will result in robustness and capability more nearly typical of biological vision. Biology has been chary with its circuit engineering. It is generally believed that cortical architecture is quite similar throughout the brain, suggesting that a common computational repertoire may be at work in a variety of perceptual, motor and cognitive functions. The problem solved by map-seeking circuits, the discovery of mappings between patterns, certainly applies to other visual problems, such as 2D image recognition and shape-from-view displacement (motion or stereo). Nevertheless there remains much work to do to accommodate the full range of biological visual capabilities with map-seeking circuits, and only time will tell whether they may be extended to such capabilities as recognition under deformation, generalization and recognition by components.

References

- [1] A. Polsky, B. Mel, J. Schiller, Computational Subunits in Thin Dendrites of Pyramidal Cells, *Nature Neuroscience* 7(6), 2004 pp 621-627
- [2] A. Angelucci, B. Levitt, E. Walton, J.M. Hupé, J. Bullier, J. Lund, Circuits for Local and Global Signal Integration in Primary Visual Cortex, *Journal of Neuroscience*, 22(19), 2002 pp 8633-8646
- [3] D. Arathorn, *Map-Seeking Circuits in Visual Cognition*, Palo Alto, Stanford University Press, 2002
- [4] D. Arathorn, Map-Seeking: Recognition Under Transformation Using A Superposition Ordering Property. *Electronics Letters* 37(3), 2001 pp164-165

[5] B.A. Olshausen, D.J. Field, Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images, *Nature*, 381, 1996 pp607-609

[6] D. Arathorn, A Solution to the Generalized Correspondence Problem Using an Ordering Property of Superpositions, submitted 2004.

3D model of T-80 and WW-2 tank courtesy 3DCafe.com