A Foveal Vision Based Approach to ATR[1]

Y. T. Zhou and R. Hecht-Nielsen
HNC Software Inc.
5930 Cornerstone Court
San Diego, CA 92121

Abstract

Unlike traditional automatic target recognition (ATR) approaches, the foveal vision based approach uses a foveated sensor to dynamically allocate visual resolution spatially and temporally to objects relevant to the task. It mimics human eye movements to find "attractive" regions quickly and accurately. A software demonstration system has been built based on the foveal vision approach and tested on a variety of sensor data including TV, infrared (IR), and laser radar (LADAR). Test results indicated that the foveal vision based approach not only supports a decrease of several orders of magnitude in the amount of data processing but also provides an increased ability to discriminate targets.

1 Introduction

Current ATR systems are highly computationally intensive and susceptible to the high-variability of target signatures and backgrounds that result from effects such as variation in illumination, aspect angles, occlusion, and obscuration. It is necessary to employ new techniques capable of managing these issues in a robust manner. Studies in the neuroscience, machine vision, and neural networks have suggested that the best way to cut the costs and increase the capabilities of machine vision systems is to move to foveated sensors. In fact, a number of impressive demonstrations of the advantages and potential capabilities of such sensors have been carried out [1, 2].

Recently, we have developed a foveal vision based ATR system under the Army VARTAC program. To derive features, current machine vision systems start with a uniformly sampled, high-resolution image, and then carry out an enormous number of arithmetic calculations. In contrast, the foveal vision based system eliminates the need for these expensive and slow intermediate steps. It measures features by foveal sampling – dynamically allocating visual resolution spatially and temporally to objects relevant to the task. Since foveal sampling uses sensor and processing resources more efficiently, it supports a decrease of several orders of magnitude in the amount of data processing needed to execute and complete a vision task. To verify the concept of this new approach, a software demonstration system has been built. The demonstration system mimics human eye movements to find "attractive" regions quickly and accurately. Test results on real data indicated that the foveal vision based approach provides an increased ability to discriminate targets.

[1]This work was supported by Army Research Laboratory under Contract DAAL01-93-C-0061.

2 The Foveal Vision Based ATR System

The foveal vision based ATR system contains five major functional modules: a foveated sensor, a saccade generator that originates movements and locates potential targets, a centering processor that stablizes movements and measures target center, a false alarm filter that eliminates false alarms, and a target classifier that generates target IDs.

2.1 Foveated Sensor

The foveated sensor measures features by foveal sampling. It is based on the modified version of Rybak's foveal rosette system [2]. The foveal rosette is a sampling frame of nodes defined by the intersections of concentric rings and radial spokes as show in Figure 1. The radius of each ring is much larger than the previous ring; this provides a nonuniform sampling pattern. At each intersection of a spoke and a ring, a set of Gabor features is gathered. The essential element of the rosette is the high density of the features near the center of the rosette and regular

falloff of features to low density at the periphery of the rosette, not their regular spacing. The specific feature set is based on the concepts of Rybak[2]. The scale of the spatial frequency features depends on its sampling position in the rosette. The scale gets smaller towards the center of the rosette to achieve a higher resolution. The point on the image that lies at the center of the rosette/foveal sampling pattern is called a *fixation point*. The movement of the rosette from one fixation point to the next, which is known as a *saccade*, is controlled by the saccade generator.

Gabor features are generated by convolving Gabor functions with the image at the rosette sampling points

$$g(i,j|m,n) = \sum_x \sum_y I(x,y)G(x-i,y-j|m,n) \quad (1)$$

where $I(x,y)$ is the intensity function of the image, $G(x,y)$ is a Gabor function, and $g(i,j|m,n)$ is the Gabor feature, (i,j) is the coordinates of the selected point, n denotes the orientation and m denotes the resolution. The orientation and resolution is determined by the modulation and scale parameters, respectively [3, 4]. Gabor features are further organized as a feature vector

$$\underline{G}(i,j) = [|g_r(i,j|m_1,n_1)|, |g_i(i,j|m_1,n_1)|, ..., |g_r(i,j|m_k,n_k)|, |g_i(i,j|m_k,n_k)|]^T \quad (2)$$

where T is the transpose operator, single vertical line $|\cdot|$ denotes the absolute value operation, and subscripts r and i denote the real and imaginary parts of the Gabor feature, respectively. Since Daugman has been shown that with a relatively small number of Gabor functions, an image/region can be vividly reconstructed from the Gabor vectors with little loss of visual details, only a few Gabor features are needed to represent a target.

A prototype demonstration model of foveated sensor which computes Gabor features has been built under the ARPA sponsored Wheel of Fortune program. Details on the demonstration model can be found in [5].

2.2 Saccade Generator

The saccade generator moves the center of the rosette to fixation points on objects of interest as shown in Figure 1. A nexus point defined as a fixation point that is located at a repeatably finable "center" of spatial frequency activity of an object. Each saccade is less than one rosette radius in length. This ensures that the image information available to the saccade generator at a given time will be sufficient.

The saccade generation is closely related to the problem of the selective attention in the human visual system. Since most information contained in the image is irrelevant to a vision task, it is necessary to focus analysis on regions of interest. Motivated by the studies of human visual behavior, the saccade generation module uses target matching patterns with a set of selection rules to determine regions of interest based on Gabor features. The selection rules which are similar to the criteria used in the human eye movement [6] are based on

- Low-level visual stimuli such as color, contrast and spatial frequency.
- High-level visual features such as vertices of polygons and axes of symmetry.
- Proximity of stimulus such as a target closer to the fovea than the others.
- High-level goal such as locating preselected targets.

For moving targets, factors such as sudden change, motion, and direction are used.

The target matching patterns have very simple geometric shapes such as rectangle, oval, etc. The basic concept is that human-made objects have geometric structure that is, at low spatial resolution, similar to such simple geometric shapes. On the other hand, backgrounds typically are not similar to such shapes, particularly when shapes of a particular scale are used. Selection of matching patterns depends on the high-level goal. Saccades have been called "programmed" to emphasize a visual target must be chosen in advance before the movement occurs. To identify points in an image which are most likely to be target locations, a similarity function is used. The similarity function compares Gabor feature vectors calculated at the rosette sampling points with a Gabor feature vector representing the target matching pattern.

The similarity function is defined as

$$S(i,j) = \frac{\underline{G}(i,j) \cdot \underline{G}(p)}{||\underline{G}(i,j)||\,||\underline{G}(p)||} \min\left(\frac{||\underline{G}(i,j)||}{||\underline{G}(p)||}, \frac{||\underline{G}(p)||}{||\underline{G}(i,j)||}\right)$$

where $\underline{G}(i,j)$ is the Gabor feature vector at point (i,j), $\underline{G}(p)$ is the Gabor feature vector of the target matching pattern, "$\cdot$" denotes the inner product, and double vertical line $||\cdot||$ is the norm operation. Since the matching pattern is fixed, its Gabor features can be precomputed and only has to be computed once.

The similarity function is normalized to $[0, 1]$. If the pattern matches the target exactly, i.e.,

$$\underline{G}(i,j) = \underline{G}(p),$$

then the similarity function gives a value of 1.0. But it is always less than 1.0 since none of targets has the same shape as the match pattern, a filled polygon. The similarity function is adopted from the one used by Buhmann *et al* [7]. To capture the local structural information, multiple feature points are used for each target. A threshold is used to determine whether the sampling point is a potential target or not. The threshold is often set at a very low level to ensure that no target will be missed.

Using the similarity measurement approach has a number of advantages. First, the detection performance in noise is good because it implements a form of matched filter. Second, using the Gabor features as reduced representation of the geometric match pattern results in a very computationally efficient implementation. Third, by using multi-spatial scales, partially occluded or obscured targets can be detected. Finally, by properly normalizing the Gabor features, contrast reversal invariance can be achieved.

2.3 Centering Processor

The centering processor stablizes the sensor and keeps the target centered after the sensor moving to the new fixation point. Due to the limited number of the rosette sampling points, new fixation points selected by the saccade generator are often off the target center. The stablization is a control problem involving an error estimator, a plant, and a control law that drives the error to zero [8]. The error signal can be derived from a variety of sources. The most important source is the vision system since the stablization depends on visual tasks. For ATR applications, the visual task is to locate targets. Therefore, the stablization can be simplified as a problem of finding the target center. The saccade generation module locates the target center based on local statistics such as local intensity distributions. Gabor features are then extracted from the target center and its surrounding points for the classification purpose. Using the foveated sensor, Gabor features can be gathered instantly once the target is centered.

2.4 False Alarm Filter and Target Classifier

Both the false alarm filter and target classifier use a multi-layer feedforward (MLF) neural network. Neural networks have been studied for many years and significant progress has been made in the past few years. With a learning algorithm [9, 10, 11], the MLF neural network is capable of partitioning the pattern space with arbitrary nonlinear decision boundaries [11]. It provides a powerful classification tool for target identification.

The false alarm filter eliminates false alarms from potential targets. Eliminating false alarms saves computation and makes the training procedure and classification task much easier for the target classifier. The input to the MLF neural network consists of multiple Gabor feature vectors which collectively represent the target. The number of neurons in the input layer is determined by the number of the feature vector components. The output layer of the MLF neural network contains multiple processing elements (PEs). The number of PEs in the output layer is equal to the total number of classes. One PE is for each possible target ID. The output is interpreted as a 1-out-of-N (for N possible classes) code, i.e., the output class is taken to be the ID associated with the PE that has the highest state. The number of the inputs and outputs can be predetermined. However, the number of hidden layers and the number of PEs in each hidden layer are variable. They depend on the properties of the features and number of the inputs and outputs, and must be experimentally determined.

3 A Demonstration System

We have built a demonstration system to verify the concept of the foveal vision based approach. As mentioned earlier, a prototype demonstration model of a foveated sensor has been built separately under the ARPA-sponsored Wheel of Fortune project. The demonstration system described in this section is a X-Windows software based simulation system. Since no foveated sensor has been used in the demonstration system, the feature extraction is carried out by the saccade generator. The demonstration system is designed to input any two-dimensional image data such as TV, IR, and LADAR and graphically demonstrate the results overlaid on the input image.

The user interacts with the system through the system's display window. As shown in Figure 3, the top part of the display window is a panel which contains seven "buttons": Load, Batch, Saccade, Center, FAF, TC, and Quit. The Saccade stands for the saccade generator, the Center for the centering processor, the FAF for the false alarm filter, and the TC for the target classifier. The Load button is used to read and

display an image and the Quit button is for closing the display window. The Batch button is mainly for running multiple saccades and the number of saccades can be defined by the user. Each button can be pressed to run the corresponding module. Pressing a button before the last module completed just queues up the request to execute that module when the last module completes. Although the typical system sequence is Saccade/Batch, Center, FAF, and TC, the user may choose to skip intermediate steps using the results of the last executed previous processing step as input to the selected processing step. The bottom part of the display window shows the currently selected image.

Figure 2 shows a second generation IR image which contains only one target near its center. The detection and classification results are given in Figure 3. A target was found immediately and classified correctly as a BMP after the first saccade. The initial fixation point was set at the image center. The plus sign "+" represents the first fixation point, the circle indicates the target center, and the number "2" overlaid on the target gives the target class. The match pattern had a rectangular shape and was used for all targets. The neural network was trained for 12 classes including clutter as given in Table 1. After 15 saccades, the whole image was completely searched and only one target was found. This performance is typical for the foveal vision based ATR system. Compared with traditional target detection algorithms/systems, the foveal vision based ATR system supports a decrease of one to two orders of magnitude in the amount of data processing required for ATR and provides an increased ability to discriminate targets.

Class	Target ID
0	Clutter
1	HUMMV
2	BMP
3	M730
4	T72
5	M35
6	ZSU23
7	2S1
8	M60
9	M113
10	M2
11	M1

Table 1: A list of target classes.

References

[1] Y. Y. Zeevi and R. Ginosar. "Neural Computers for Foveating Vision Systems". In R. Eckmiller, editor, *Advanced Neural Computers*, pp. 323–330. Elsevier Science Publishers B.V., North-Holland, 1990.

[2] I. A. Rybak, N. A. Shevtsova, L. N. Podladchikova, and A. V. Golovan. "A Visual Cortex Domain Model and Its Use for Visual Information Processing". *Neural Networks*, vol. 4, pp. 3–13, 1991.

[3] Y. T. Zhou and R. Crawshaw. "Contrast, Size and Orientation Invariant Target Detection in Infrared Imagery". In *Applications of Automatic Object Recognition*, SPIE Proc., vol. 1471, pp. 404–411, Orlando, Florida, April 1991.

[4] Y. T. Zhou and R. Chellappa. *Artificial Neural Networks for Computer Vision*. Springer-Verlag, New York, 1992.

[5] R. Hecht-Nielsen and Y. T. Zhou. "A Low Cost Foveal Vision System". In *Proc. Goverment Neural Network Applications Workshop*, Dayton, OH, August 1992.

[6] A. L. Abbott. "A Survay of Selective Fixation Control for Machine Vision". *IEEE Control System Magazine*, pp. 25–31, Aug. 1992.

[7] J. Buhmann, J. Lange, and C. von der Malsburg. "Distortion Invariant Object Recognition by Matching Hierarchically Labeled Graphs". In *Proc. Intl. Joint Conf. on Neural Networks*, vol. I, pp. 155–159, Washington, D.C., June 1989.

[8] M. J. Swain and M. A. Stricker. "Promising Direction in Active Vision". *International Journal of Computer Vision*, vol. 11, pp. 109–126, 1993.

[9] D. B. Parker. Learning-Logic. Technical Report TR-47, Center for Computational Research in Economics and Management Science, 1985.

[10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, vol. 1, pp. 318–362. MIT Press, 1986.

[11] J. Makhoul, A. El-Jaroudi, and R. Schwartz. "Formation of Disconnected Decision Regions with a Single Hidden Layer". In *Proc. Intl. Joint Conf. on Neural Networks*, vol. I, pp. 455–460, Washington, D.C., June 1989.

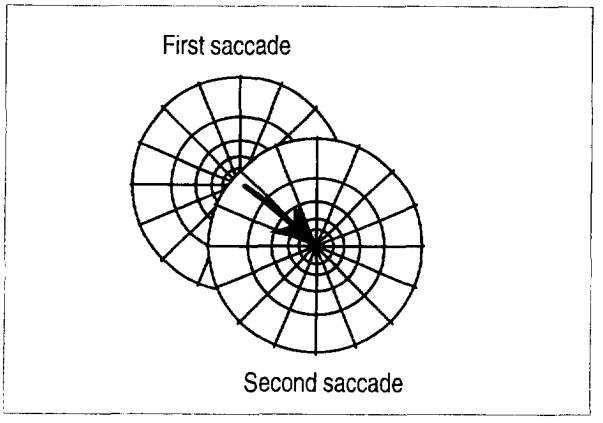

Figure 1: The saccade generator uses the features gathered at the current fixation point to generate a saccade. The arrow indicates the moving direction from the first fixation point to the second fixation point. The foveal rosette has 5 rings and 16 spokes.

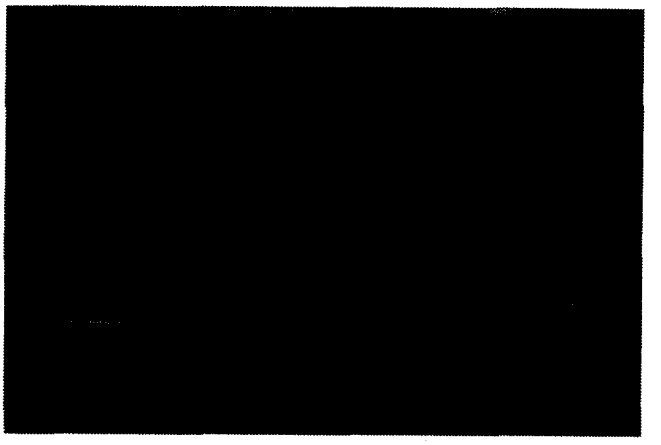

Figure 2: A second generation IR image which contains only one target near its center.

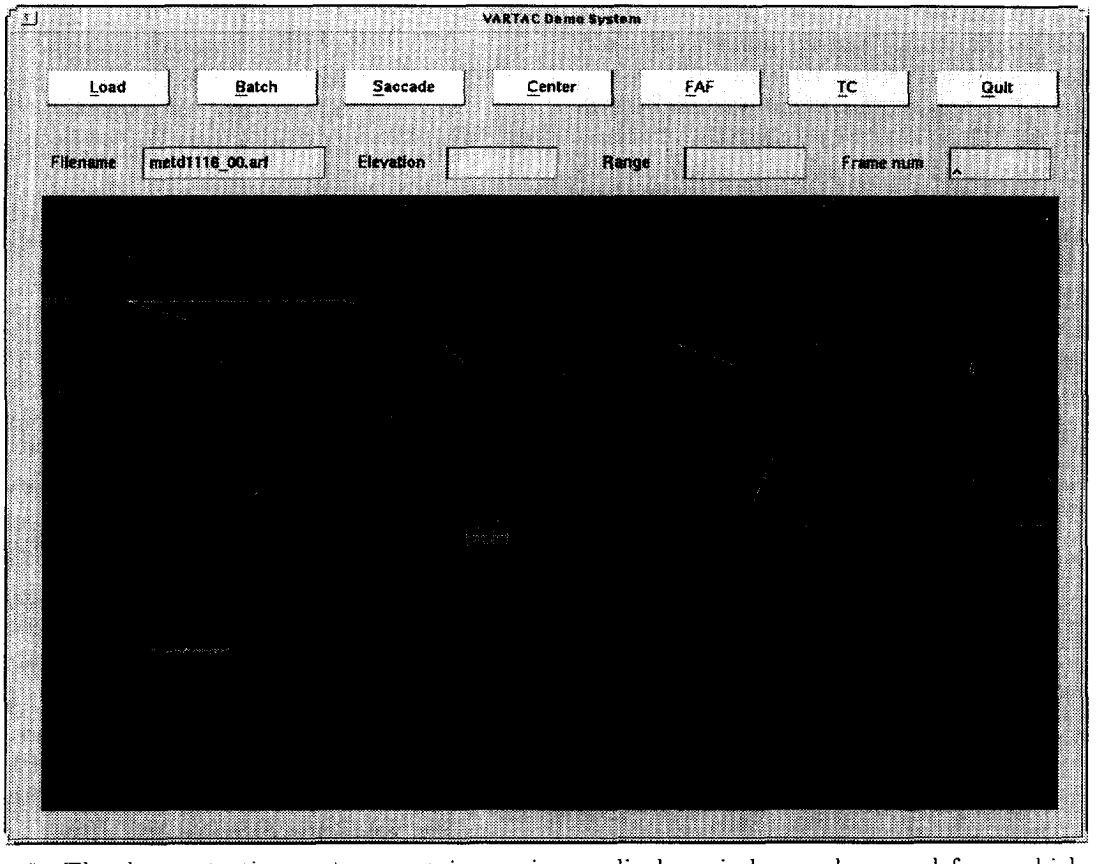

Figure 3: The demonstration system contains an image display window and a panel from which seven different modules can be launched. A target was successfully detected and correctly recognized after the first saccade. The initial fixation point was set at the image center. The cross sign "+" represents the first fixation point, the dark circle indicates where the target center is located, and the number "2" overlaid on the target gives the target class. A list of 12 target classes including clutter is placed on the upper left corner of the image. After 15 saccades, the whole image was completely searched and only one target was found.