Automatic Target Recognition: The Problems of Data Separability and Decision Making

E. J. Hughes, M. Lewis
C. M. Alabaster
Cranfield University at The Defence Academy of the UK
Shrivenham, Swindon, SN6 8LA
Lt., F. Soldani, ITAF
ejhughes@theiet.org, m.lewis@computer.org, c.m.alabaster@cranfield.ac.uk

This paper treats the problem of target recognition as a decision process. The nature of the decision to be made has a direct bearing on the data gathered and the subsequent processing. A key factor in the processing is the separability, i.e., the ability to distinguish, of radar images of similar but distinct objects. A number of recognition algorithms are considered and their suitability for data sets of various types is discussed. In addition some simple measurements of the transfer functions of two targets are considered. Observation suggests that the examples have characteristics that may make them readily separable. As with all recognition techniques the quality and quantity of training data available will place a limit on the performance of any recognition technique and this is discussed in the text. The view is formed that a single technique is unlikely to be successful and several techniques cued by gross-features of the image may be more appropriate.

Introduction

Accurate and reliable target recognition is of critical importance in many military radar applications as well as a significant number of civilian surveillance scenarios. The consequences of erroneous recognition can range from an inconvenience to catastrophic incidents involving unintended loss of life. Unfortunately, robust target recognition is a complex and challenging task and further work across numerous radar application environments is required.

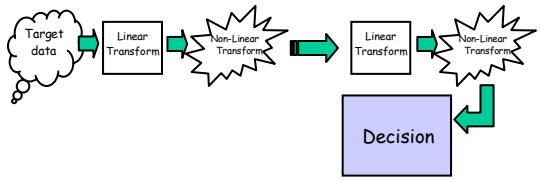

Figure 1 Target Identification Abstract Framework

The target identification process is a chain of linear and non-linear processes, as illustrated in Figure 1, resulting in a decision (or sequence of decisions) that allow the target classification to be deduced. Care must be taken in the design of the processing chain to exploit information content to the maximum in order to make the most informed decision possible.

The aim-point of current work at Cranfield is to develop techniques of information analysis to assess the performance bounds of target recognition.

Decision Making and Separability

The fundamental reason for gathering data is to make a decision. The decision should determine what data is gathered, and to what precision. Therefore the fundamental Automatic Target Recognition limitation is: What question is to be answered and is the information content of the data available sufficient?

Each individual decision is a binary process. A decision is easiest to make if the supplied information is unambiguous – e.g. 99:1 or 51:49. Can the decision be made with confidence, given the information available?

The decision is made by assessing a *feature vector* which forms a concise summary of the observed data points from the target. The feature vector is often multi-dimensional.

For a decision making process, often the decision based on the observation of a single feature vector is considered. If the final decision can be postponed until after multiple observations, either all of the data may be considered as one large feature vector, or the feature vectors may be smoothed (integrated) before the decision, or a binary integration of the decisions may be used to improve the observation. There is no reason why feature vectors with different structures and decision boundaries may not be gathered at each new time instant, based on the prior decisions made. This process would lead to a tree

structure of decisions. Structures of this type or similar are often applied in expert systems and Bayesian belief networks [1, 2]

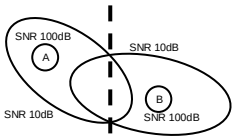

Figure 2. Separability of Two Classes

Figure 2 shows two target classes, A and B, that have been assessed using a 2-dimensional metric vector. The ellipses bound all of the feature vectors that have been observed from a wide range of targets from classes A and B. The classes are shown separated by a linear division plane (shown as a dotted line) at high signal to noise ratios. However, the superimposed class regions at a lower SNR are only partially separable, with some ambiguity existing. The linear decision plane shown is still the optimal decision solution, even though a definitive decision can never be made, given only one observation of the two-dimensional feature vector. It is also clear that the feature on the Y axis contributes very little as it is parallel to the decision plane. All of the classification is performed by the feature on the X axis.

If the classes are not totally separable, it is not possible to guarantee a clear decision: the decision becomes probabilistic, each class has a probability of occurring, given the feature vector.

This argues for more complex decision boundaries giving A, B and indeterminate decisions. The latter may be used to cue further testing.

For a 'clean' decision, data points MUST be linearly separable, therefore data transforms may be required. This raises two key questions:

- How separable are the different classes and how will the separability degrade with measurement noise?
- How sensitive are the separability transforms to different elements of the feature vector being obscured by clutter?

A classic example is the detection of targets in coherent radar data. At each range cell, the in-phase (I) and quadrature (Q) voltages are measured and a decision must be made as to whether there is a target present or only noise.

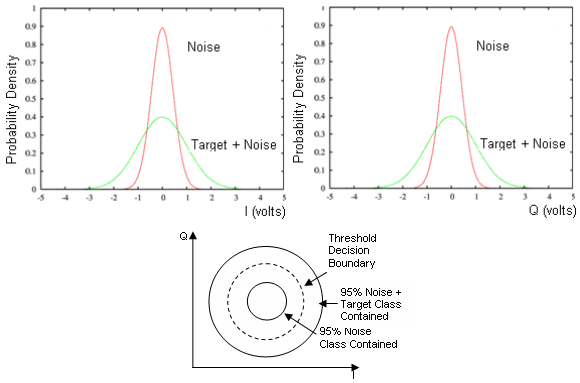

Figure 3. Cartesian Representations

Figure 3 shows the probability density functions for the I and Q voltages for the noise and target-plus-noise classes. It is clear that both noise and target-plus-noise have near-Gaussian density functions and overlap almost entirely being only partially separable in the very tails of the distributions. Classification is possible, but when plotted in the 2D plane, the threshold is a circular region.

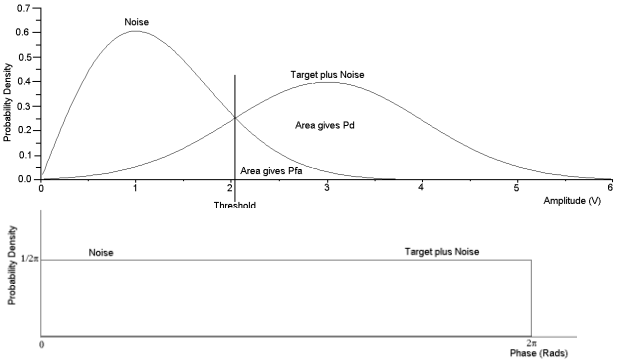

Figure 4. Polar Representations

In contrast, after a Cartesian-to-Polar transformation (Magnitude and Phase), the probability density functions shown in Figure 4 are obtained.

It can be seen that the probability density function of the phase component is uniform and identical for both noise and target-plus-noise and thus they are totally inseparable on the basis of phase. However, the probability density functions of the magnitude are partially separable with a linear decision plane. This is an example of transforming a convexly

separable problem to linear separability before the decision making step.

The probabilities of correct and false classification describe the degree of separability possible with the decision boundaries employed. If the appropriate decision boundaries are chosen (circular for Cartesian, linear for polar), the same classification results can be achieved for the same input data. In practice however, either the noise or target returns (or both) may not be truly Gaussian in nature, and therefore after the polar transform, the phase component may not have a true uniform distribution and the polar transform approach will result in the rejection of some information as the phase is ignored in the decision making process.

Transformations

Tools exist that can help identify appropriate linear transformations. Non-linear transformations are more difficult. Often the classes remain only partially separable after the transformations.

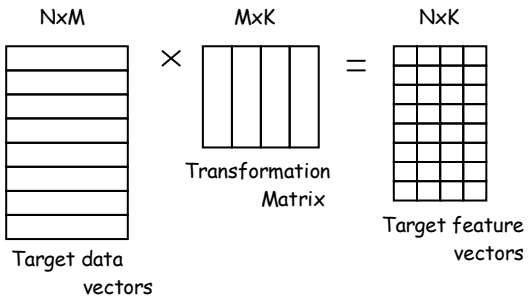

Figure 5. Linear Transformation

Any linear transformation can be represented as a matrix multiplication. The generalised linear transform is shown in Figure 5. A sequence of linear transformations can be represented as a single transformation.

The dimensionality of the metric vectors can be increased or decreased ($K>M$, $K=M$, $K<M$). Ultimately K must equal the number of target classes. If the rank of the transformation matrix is less than M, then information is lost (the transform is not invertible).

Artificial Neural Networks

As an example of a non-linear transformation system that can have its decision boundary adapted based on examples of feature vectors that are used for training, Figure 6 shows an artificial neural network or perceptron. The perceptron uses a hard threshold in each neuron and provides a classification decision.

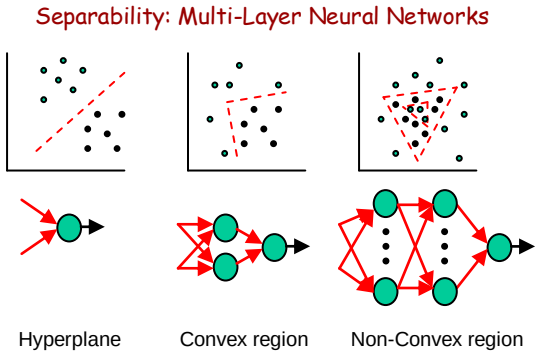

Figure 6. Separability characteristics for Various ANN

Three different network structures are shown, along with the types of decision boundaries they can generate. A single neuron provides a single linear separation plane and can only perform classification on problems with sufficient linear separability. With one hidden layer, each neuron in the hidden layer can provide one linear separation plane. The output neuron can then combine these linear planes (effectively a logical operation) to form a convex region and can handle linear or convexly separable data. With two hidden layers, the first layer forms linear planes, the second layer then forms multiple convex regions based on the linear planes, and the output neuron can then form logical operations on the multiple convex regions. As both union and intersection are valid, then multiple convex regions, non-convex regions, or regions with 'holes' in can be generated.

The complexity of the neural network is the main factor in determining how well it will perform in a given situation. With too few neurons in the layers, the network lacks the ability to form the complex structures it requires to partition the data set. With more layers, the partitioning can become more complex, but at the expense of far more weights.

If too many neurons are used in a hidden layer, the network may learn the training data exactly, but may not generalize (ie. The decision boundary fits the contours of the training data exactly) and therefore work only on the test data. If too many layers are

used (allowing a very convoluted decision boundary), the network may learn to respond to noise within the training set, and again not provide a generalised response.

The Venn diagram like form of Figure 2 illustrates a further point. The modern form of the Venn diagram has an outer boundary signifying that the universe of discourse is finite. It may be assumed that since the magnitudes of the features are also finite that attempts to increase the number of classes will reduce the number of disjoint spaces due to overlap. Increasing the number of classes will reduce the separability.

Empirical work [3] has shown that the best performance of a neural network occurs when the number of hidden nodes is equal to $log_2(T)$, where T is the number of training samples. This value represents the optimal performance of the neural network as well as the optimal associated computational cost. It has also been shown that the number of training patterns must be greater than the total number of separable regions in the input space [4]. In addition, in a d-dimensional space, the maximum number of regions that are linearly separable using $\boldsymbol{H}$ hidden nodes is given by

$$M(H,d) = \sum_{k=0}^{d} \binom{H}{k}, \qquad H \geq k$$

Practical Aspects of Target Features

The above presented theory assumes that the feature sets of differing targets are disjoint or nearly so. In practice this is not necessarily the case.

Type	Feature	Description
Radiometric	*MEAN*	mean intensity
	CVAR	coefficient of variation
	WFR	weighted rank fill ratio
Geometric	*AREA*	area of target
	NN	neighbour number
	LAC	lacunarity index
	LEN	length of target
	WID	Width of target
Polarimetric	*VVVH*	polarimetric (VV/VH) power ratio

Table 1. Explanation of Features

Van den Broek and Dekker [5] have taken ISAR data derived from 3 targets in a total of 16 configurations at various aspect angles. A total of 11 features, Table 1, were extracted as pdfs from the data and the Kolmogorov-Smirnov distance between their cumulative distributions was calculated. This is illustrated in Figure 7 where the pdf of the 'Area' feature is shown.

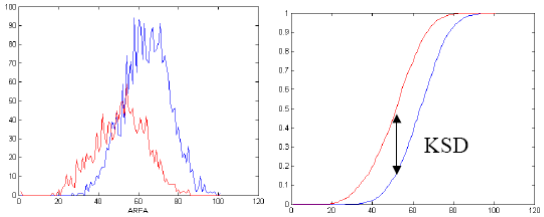

Figure 7. Kolmogorov-Smirnov Distance

Feature	Average intra-class	Average Inter-class
MEAN	0.09	0.29
CVAR	0.06	0.16
WFR	0.05	0.19
AREA	0.12	0.41
NN	0.09	0.38
LAC	0.09	0.39
LEN	0.08	0.39
WID	0.16	0.22
VVVH	0.05	0.14

Table 2. Average Kolmogorov-Smirnov Distances

The Kolmogorov-Smirnov distance is the maximum distance between the cumulative distribution functions of two pdfs. A value of zero implies that the two distributions are identical, a value of one implies complete separability. Table 2 shows the mean results. As might be expected the distances for intra-class objects are small. Unfortunately the interclass distances are less than 0.5 indicating that the objects may not be easily separable on the basis of this feature set.

Figure 2 illustrated the separability problem. The Kolmogorov-Smirnov (KS) distance is a practical indicator of separability.

It may be commented that the features cited here are not necessarily unique to a given object and many do not capture shape or structure as would be used by a human operator.

Hidden Markov Models

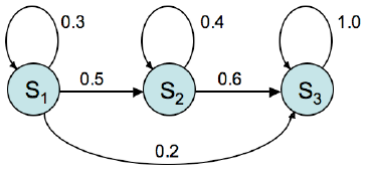

Figure 8. Markov Chain

The constituent element of a Hidden Markov Model (HMM) is the Markov chain, Figure 8, and can be used to model a sequential signal. The circles in this figure represent the discrete states, i.e., sub-elements, of the modelled signal and are associated with one or more time steps of the input.

The model moves from one state to another at regular discrete intervals equal to the time interval of the processing. The movement is random and is described by the transition probabilities assigned to each path option leaving the current state. The signal may also remain in a given state and this is represented by the loops in the diagram. The next state to be entered is dependent only upon the current model state and is not influenced by the sequence of states passed through to reach the current state. This process is hence referred to as a first order Markov process.

When the model is in a given state, it generates an output or observation, which will be a feature vector for a target recognition implementation. The particular feature vectors output by the model are governed by a distribution which gives the probability that any set of feature values will be generated when the model is in the associated state. Only the feature vectors being output by the model are observable and the sequence of state transitions are hidden from the observer, hence the name Hidden Markov Model.

HMM Probability Calculations

Suppose that we have a set T of targets and a separate training set for each target. An HMM is built for each target using the associated training set.

The problem is to calculate the relative likelihood of each model emitting the observed sequence of feature vectors. If the HM model associated with the target, τ, has parameters, λ_τ then when presented with a sequence of observations, σ, choose the target with the most likely model, i.e.,

$$\tau^* = \max_{\{\tau \text{ in } T\}} \left(p(\sigma \mid \lambda_\tau) \right)$$

The model output for a complete target is the result of N state transitions and is equal to the number of feature vectors which represent the target. If the identifier for the current model state is σ_m, where m denotes the current time step and the start state is B and the end state E, which are not associated with any feature vector, then transition probabilities from B to each permitted start state can be specified as well as transition probabilities from each permitted last state to the end of the target. Hence the probability of the model emitting the complete sequence is:

$$p(\sigma_1, \sigma_2, \ldots, \sigma_N)$$
$$= t_{Bs_1} \cdot \sum_{\substack{\text{over all} \\ \text{sequences} \\ \text{of length} N}} \left(\prod_{m=1}^{N-1} t_{s_m s_{m+1}} \cdot p(\sigma_m \mid s_m) \right) \cdot p(\sigma_N \mid s_N) \cdot t_{s_N E}$$

In the above, σ_m is the m^{th} element of the sequence σ, $t_{s_m s_{m+1}}$ is the probability of a transition from state s_m to s_{m+1} and $p(\sigma_m|s_m)$ is the probability that the state s_m could produce the feature vector σ_m.

The preceding equation gives the probability of the model generating the observed features, taking into account all possible sequences of states. Determination of the relative likelihood of the target represented by the model being present can then be calculated using Bayes' theorem.

$$p(\lambda_\tau \mid \sigma) = \frac{p(\sigma \mid \lambda_\tau).p(\lambda_\tau)}{p(\sigma)}$$

$p(\sigma)$ does not depend upon the model being considered and can be ignored. In the recognition scheme, if the occurrence of all targets is equally likely, then $p(\tau)$ may also be ignored and identifying the target model that maximises $p(\sigma|\lambda_\tau)$ is sufficient.

HMM Training

The training problem is determining how the parameters of the model (the probability distributions for the transitions and feature vector

outputs) can be adjusted to maximise the probability that the model created the output sequence. Training of a HMM is of critical importance to the recognition abilities of the model and requires a significant amount of training data. Additionally, in order to gain optimal performance, the statistics of the training data must be representative of the target samples which will be provided as input during recognition.

Gaussian Mixture Models

In many HMM implementations the probability distribution associated with each state is provided by Gaussian Mixture Models (GMM) [6]. A Gaussian mixture is a weighted sum of Gaussian densities as illustrated in Figure 9 for a mixture of three single Gaussians.

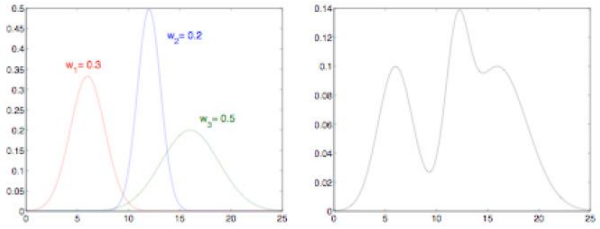

Figure 9. Gaussian mixture composed of 3 Gaussian pdfs

Training a GMM to model a complex pdf requires adjustment of the number of Gaussians, the weights and the means and covariances of each Gaussian pdf in an iterative manner until an optimal fit of the GMM to the training data is achieved. This training normally uses the expectation-maximization (EM) algorithm [7] ans yields a Maximum Likelihood (ML) estimate of the distribution parameters.

Applying HMMs to Target Recognition

HMMs are particularly suited to the processing of sequential data. As such they have found applications in speech processing and face recognition. They may also be applied to radar imaging if data ordering is used to convert the 2D spatial image representation into a sequential representation. Such a representation avoids the need for fully connected 2D HM models, which are exponentially complex in the size of the image and require intensive amounts of training data [8].

Target Identification and Optimised Illumination

Bell [9] has investigated the optimisation of radar target illumination from an information- theoretic viewpoint for both target detection and target identification. The key concept is that of the target impulse response, the Fourier Transform of the target frequency response for all frequencies. .

The conclusion of the work was that for optimum detection radar power should be concentrated at those frequencies which gave the greatest return whilst for optimum information for identification greater power should be transmitted at those frequencies where the return is weak than where the return is strong. The argument is that information is contained in these areas and by raising the signal power a signal to noise ratio approaching that of the strong areas can be obtained with a consequent reduction in uncertainty.

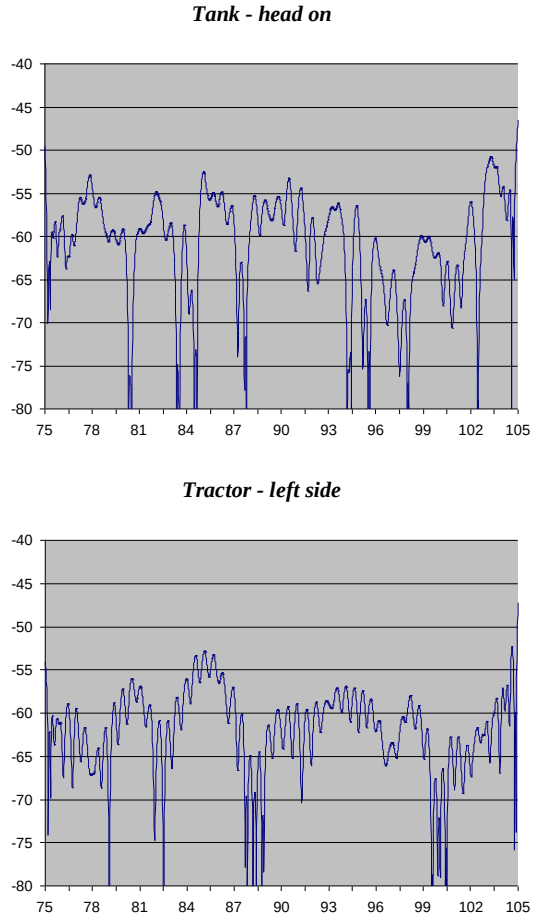

Figure 10. Transfer Functions for Test Targets

A recent Cranfield MSc., project investigated matched illumination and contains information that may be usefully applied to target identification [10].

The project used a Vector Network Analyser (VNA), in the reflection coefficient mode to measure the frequency response of two different types of test target. The targets were 1/32nd scale models of a main battle tank and a farm tractor and were measured in the 90GHz region. Target classification was not the aim but the difference between the transfer functions, particularly in the zeroes which are strongly evident in Figure 10, could represent a first step in the classification of a target. Although only two widely differing targets have been examined the results suggest that a wider investigation would be worthwhile.

The measured results were used in a matched illumination simulation. Since it is not easily feasible to amplitude modulate a pulsed radar transmission to produce the required matched illumination a 'matched energy' approach was used. A constant amplitude frequency sweep with a variable rate such that a long dwell occurred on those frequencies that gave the strongest returns was used. Improvements in the return amplitude compared with a simple pulse of up to 20dB were observed.

The optimum waveform for identification could be obtained in a similar manner in that the sweep rate could be reduced over the low amplitude return frequencies. The action would be analogous to reducing the transmission rate in a noisy channel to obtain a 'clean' transmission in accordance with the Shannon-Hartley Theorem.

Concluding Remarks

No one method of identification will be adequate under all circumstances and for all targets. Bespoke feature vectors are necessary, and the feature vector may need to be altered, depending on the anticipated target set.

No one type of sensor will be adequate under all circumstances, if a sensor cannot provide data such that the target set appears separable, then no reliable classification can be made

Rules for method/sensor cueing are required to allow the best sensor to be directed at the target set. It may even require man-in-the-loop as one of the sensors if human identification cannot be surpassed under certain conditions.

The impact of noise/clutter degradation on separability must be established for any fielded system to ensure that accurate misclassification probability can be controlled.

Bounds, analogous to Cramer-Rao, on target separability need to be established to allow algorithms to be compared to a realistic reference.

References

1. Tsang, C., Woo, M., Bloor, C., *An Object Oriented Intelligent Tourist Advisor System*, Proc., Australian and New Zealand Conf., on Intelligent Information Systems, pp.6-9, 1996
2. Schuller, B., Muller, R., Rigoll, G., Lang, M., *Applying Bayesian Belief Networks in Approximate String Matching for Robust Keyword-Based Retrieval*, Proc., ICME 2004, Vol., 3, June 2004, pp., 1999-2002
3. Wanas, N., et al., *On the Optimal Number of Hidden Nodes in a Neural Network*, Proc., IEEE Canadian Conf., on Elec., and Computer Eng., 1998, Vol., 2, pp., 918 – 920
4. Mirchandani, G., Cao, W., *On Hidden Nodes for Neural Nets*, IEEE Trans., on Circuits and Systems, Vol., 36 No., 5, May 1989
5. van den Broek, A., Dekker, R., *Target discrimination in polarimetric ISAR data using robust feature vectors*, SET-096 / MATRIX 2005, Robust Acquisition of Relocatable Targets using MMW Sensors, NATO School, Oberammergau, May 2005
6. Day, N., *Estimating the Components of a Mixture of Normal Distributions*, Biometrika, Vol., 56, No. 3, Dec., 1969, pp. 463-474
7. Dempster, A., Laird, N., Rubin, D., *Maximum Likelihood from Incomplete Data via the EM Algorithm*, J. Royal Statistical Society, Series B, Vol., 39, No., 1, 1977, pp. 1-38
8. Kottke, D., Fwu, J., Brown, K., *Hidden Markov Modeling for Automatic Target Recognition*, 31st Asilomar Conf., on Dignals, Systems and Computers, Vol., 1, 1997, pp. 859 – 863
9. Bell, M., *Information Theory and Radar: Mutual Information and the Design and Analysis of Radar Waveforms and Systems*, PhD Thesis, CIT, Pasadena, Calif., 1988
10. Soldani, F., *Matched Illumination*, MSc., Thesis, Cranfield University, 2006